STATISTICAL MODELLING OF LONGITUDINAL LUNG FUNCTION DATA

A Thesis submitted to the
College of Graduate Studies and Research
in Partial Fulfilment of the requirements
for the Degree of

Doctor of Philosophy

in the College of Medicine
University of Saskatchewan
Saskatoon

By

Punam Pahwa

Permission to use

In presenting this thesis in partial fulfillment of the requirements for a Postgraduate degree from the University of Saskatchewan, I agree that the Libraries of this University may make it freely available for inspection. I further agree that permission for copying of this thesis in any manner, in whole or in part, for scholarly purposes may be granted by the professor or professors who supervised my thesis work or, in their absence, by the Head of the Department or the Dean of the College in which my thesis work was done. It is understood that any copying or publication or use of this thesis or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to the University of Saskatchewan in any scholarly use which may be made of any material in my thesis.

Requests for permission to copy or to make other use of material in this thesis in whole or part should be addressed to:

Director, Centre for Agricultural Medicine
College of Medicine
University of Saskatchewan

ABSTRACT

Statistical models were developed for the analysis of longitudinal data obtained from two different studies. For the first longitudinal study, our objective was to evaluate different longitudinal models for predicting the longitudinal decline in lung function measures of grain elevator workers and to assess the goodness of fit of these models. Generalized estimating equations and maximum likelihood methods were used to fit different models, assuming data were equally spaced in time. Concordance coefficients r_c and $r(\hat{\omega})$ were used to assess the adequacy of the model and variance-covariance structure respectively. Pseudo-likelihood ratio test, $\hat{\lambda}$, was used to test the null hypothesis that the assumed covariance structure is equal to the true covariance structure. To predict the annual decline in population pulmonary function, a transitional model proved to be the most parsimonious. To predict the annual decline in an individual grain workers, a random effects model with compound symmetric covariance structure gave a better fit. The values of pseudo likelihood ratio test depend on the sample size and the covariance structure. Concordance coefficient $r(\hat{\omega})$ has limited range of values when sample size gets large. Random effects models were also fitted assuming observations were unequally spaced in time. An important finding from the random effects models was that there might be more observational error in measuring FVC than FEV_1.

Another longitudinal analysis was conducted to study the respiratory health effects of initial exposure to grain dust among workers commencing employment in the grain industry in the Province of Saskatchewan. In this study, we examined the factors associated with bronchial hyperresponsiveness among grain elevator workers. We used

correlated survival data analysis techniques to determine predictors of bronchial hyperresponsiveness. Consistent estimates of standard errors were obatined by using jackknife, bootstrap and the method proposed by Wei, Lin and Weissfeld (WLW). We conclude that survival analysis is a useful technique to analyze the bronchial hyperresponsiveness data. The estimates of standard errors were very similar for jackknife, bootstrap, and WLW, but different from those obtained using standard likelihood maximum methods.

Cox's proportional hazard model based on the data from first longitudinal study, proved to be a useful technique in investigating the relationship between survival time (time to first episode of wheezing) and possible prognostic variables.

Acknowledgements

I wish to express my sincere appreciation to my supervisor Dr. A. Senthilselvan for his continued inspiration and advice for high and better work throughout the course of my Ph.D and preparing this thesis. I wish to thank each of my committee members, Dr. D. Sherrill, Dr. M. Bickis, Dr. D. Cockcroft, and Dr. L. Tan for their suggestions and comments during this process.

I sincerely thank Dr. James Dosman for providing me with my own office and a computer during this training. A special thanks goes to Dr. H. McDuffie for her moral support and proof reading help.

I sincerely thank the Centre for Agricultural Medicine for providing me with the databases on which this thesis was based. For my financial support during the course of Ph.D, I am grateful to College of Medicine, College of Graduate Studies and Research, and Founding chairs - Centre for Agricultural Medicine.

Finally a very special thank you to my family: my husband Dr. Ranvir Pahwa, my son DaveRishi and my daughter Divya (sometimes who wondered if life will ever return to normal) for their continued love and support during my Ph.D. Therefore this thesis is dedicated to these people.

TABLE OF CONTENTS

LIST OF TABLES

List of figures

ABBREVIATIONS

ANOVA- Analysis of Variance

AR - Autoregressive

BAN - Best Asymptotically Normal

BLUE - Best Linear Unbiased Estimator

BLUP - Best Linear Unbiased Predictor

CTS - Carpel Tunnel Syndrome

EM - Expectation-Maximization

GEE - Generalized Estimating Equations

FDM - First Difference Model

FEV_1 - Forced Expiratory Volume in one second

FVC - Forced Vital Capacity

GDMSP - Grain Dust Medical Surveillance Program

GLM - Generalized Linear Model

LR - Likelihood Ratio

MANOVA - Multivariate Analysis of variance

ML - Maximum Likelihood

NGW - New Grain Workers'

NR - Newton Raphson

REML - Restricted Maximum Likelihood

PC_{20} - Provocation Concentration

PD_{20} - Provocation Dose

1. INTRODUCTION

1.1 Models for Longitudinal Data Analysis

In cross-sectional data, only a single response y_i (i = 1, 2, ... , n) and a px1 vector $\underline{x}$ of covariates are available for each of the n experimental units. In contrast to cross-sectional data, longitudinal data are comprised of repeated observations y_{it} $(i = 1, 2, \ldots, n; t = 1, 2, \ldots, k)$ and px1 vector $\underline{x}_{it}$ over time t for each of the i^{th} experimental units. Missing observations over time for an experimental unit arise quite often in longitudinal studies and lead to unbalanced patterns of observations. These missing data can arise purely by chance, may be due to ethical issues related to the study or to voluntary dropout by a study subject (e.g. in epidemiological studies) due to health reasons. Also, the repeated observations over time for each experimental unit tend to be correlated with one another. Depending on the magnitude of within-subject correlation, one may need to account for this correlation when conducting statistical analysis of longitudinal data. Longitudinal studies have several challenges/complexities because of unbalanced designs, correlated observations, and time-varying covariates. These complexities make some of the traditional statistical analysis procedures, such as repeated measures ANOVA (1971) and growth curve models [Wishart , 1938; Box, 1950; Rao, 1958; Rao, 1959; Leech & Healy, 1959; Rao, 1961; Healy, 1961; Elston and Grizzle, 1962; Elston, 1964; Potthoff and Roy, 1964; Rao, 1965; and Grizzle and Allen, 1969] inapplicable.

For Gaussian data, longitudinal analysis methods are well developed, e.g. the random effects model (REM) for unbalanced longitudinal data (Laird and Ware, 1982). Estimation of the parameters of a random-effects linear model is based on iterative maximum likelihood or restricted maximum likelihood methods. The estimation of the parameters can be computationally problematic, especially when the data set is quite large and unbalanced. For non-gaussian data, extra information in addition to the first two moments is needed to determine the likelihood. This additional information leads to too many nuisance parameters, and estimation of the parameters using maximum likelihood methods becomes very difficult (Liang and Zeger, 1986). The generalized estimating equations (GEE) approach re-invented by Liang and Zeger (1986) is based on the multivariate quasi-likelihood theory, which can handle the complexities of longitudinal studies. In longitudinal studies four type of responses are encountered: continuous, discrete, count and survival. In this thesis, longitudinal statistical analysis techniques were utilized to analyze continuous, discrete (binary) and survival type data.

One of the data sets used in this thesis was collected as part of the Labour Canada National Grain Dust Medical Surveillance Program between 1978 and 1993 (detailed explanation of the data set is in Chapter 3, Section 3.2). This national program was commenced to monitor the effect of grain dust on lung function among grain elevator workers on a continuing basis. Descriptive statistics of these data sets have been reported in Technical Reports (Labour Canada, 1984; Dosman et al., 1987; McDuffie et al., 1989; and Pahwa et al., 1994). Results of the longitudinal analysis of the data consisting of the 2nd and 3rd observations were reported by McDuffie et al (1991). Longitudinal estimates of pulmonary function decline over three observations

were reported by Pahwa et al (1994). The complete data set over five observations has never been analyzed to study the long-term effect of grain dust exposure and the annual decline in lung function measurements. One possible way to analyze the longitudinal lung function data is to utilize widely applicable GEE methodology (Liang and Zeger, 1986). The GEE approach has been used to study the long-term effect of asbestos exposure (Glencross et al., 1997) and cotton dust (Christiani et al., 1999) on lung function. Only a few longitudinal studies have attempted to study the relationship between the long-term effect of grain dust and decline in lung function (Tabona et al., 1984; Broder et al., 1985; Huy et al., 1991; McDuffie et al., 1991; Zejda et al., 1992; and Pahwa et al., 1994). This thesis describes the use of different statistical methods based on maximum likelihood/restrictive maximum likelihood (random effects models) and multivariate quasi-likelihood (marginal and transitional models) to evaluate the relationship between long-term effect of grain dust and decline in lung function.

1.2 Goodness of fit

An important part of any model selection process is the assessment of how well the model fits the data (goodness-of-fit). Goodness of fit can be defined as the degree to which a predicted value from the model ($\hat{y}_i$) agrees with an observed value (y_i). In longitudinal data analysis, to date much of the work has focused on various methods for estimating and comparing the parameters of different models. Very little work has been done in the area of model selection and goodness-of-fit. To date most common approaches to assessing goodness-of-fit in longitudinal data analysis are likelihood based measures, such as Akaike's information criterion or likelihood ratio test for nested

models. There are some drawbacks to likelihood based measures of goodness-of-fit statistics. They require complete specification of the likelihood function and repeated fittings of the data to a family of nested models. In last few years, a group of investigators have provided a goodness-of-fit statistics [Vonesh, 1992 and Vonesh et al 1996] based exclusively on the model at hand. These statistics have not been used extensively to evaluate how well these statistics perform utilizing various longitudinal data sets from different scientific fields. In this thesis computer programs were written for these recently developed goodness-of-fit statistics and applied to the longitudinal lung function data to assess the adequacy of models and variance-covariance structures.

1.3 Cox's Proportional Hazard Model for survival data

1.3.1 Uncorrelated Survival Data

Techniques developed for the analysis of survival data can also be used in the analysis of longitudinal data. In longitudinal studies uncorrelated survival data can arise, when time to the first occurrence of a disease is of interest and the additional events which occur after the first event are ignored.

In GDMSP data, one interest was to determine predictors of the first episode of wheezing. It has been shown that the onset of wheeze and dyspnea was significantly related to the annual rate of decline in forced expiratory volume in the first second among non- asthmatic subjects (Jaakkola et al, 1993). Recurrent episodes of wheeze and dyspnea were related to asthma (Salvaggio et al., 1986). It has also been shown that wheezing is associated with bronchial hyperresponsiveness (Woolcock et al., 1987; Rijcken et al., 1987; Sparrow et al., 1987; Dales, 1987; Burney et al., 1989; and Trigg et

al, 1990). To determine the predictors of time to first episode of wheezing among Canadian grain elevator workers, Cox's proportional hazard model was used in this thesis.

1.3.2 Correlated Survival Data

Correlated censored survival data are observed in several biomedical studies in which two or more distinct events or failures occur for each subject during the continuous follow-up of subjects (Wei et al, 1989; Lee et al, 1992; Lipsitz et al, 1994; Lipsitz and Parzen, 1996). The failures may be repetitions of same kind of event during the follow-up. In such longitudinal studies, an important assumption of mutually independent failure times in Cox's proportional hazard model is violated. One approach to account for the dependencies between failure times is given by Wei et al. (1989). In this approach the covariance matrix of regression estimates is adjusted to account for the within-subject correlation, and are known as "variance-corrected" models (Therneau, 1997). As a part of New Grain workers' study (explained in Chapter 3, Section 3.3), correlated bronchial hyperresponsiveness data were available to us. To our knowledge, survival analysis techniques for correlated data have not been previously used for the analysis of bronchial hyperresponsiveness data. In this thesis, Cox's model was used to obtain the regression estimators and three different methods were used and compared to obtain the covariance-matrix adjusted for correlation for these regression estimators

1.4 Objectives

Two longitudinal data sets on grain elevator workers were available at the Centre for Agricultural Medicine, University of Saskatchewan. Analyses of these data sets were conducted to i) explore the role played by different longitudinal models in predicting the longitudinal decline in lung function measures of grain elevator workers; ii) assess the goodness of fit of the above models; iii) determine the predictors for first episode of wheezing; and iv) apply censored survival data techniques for correlated data to determine the predictors of airway hyperresponsiveness.

2. REVIEW OF LITERATURE

2.1 Models for longitudinal data analysis

Common characteristics of longitudinal studies are: (i) correlated responses; (ii) observations taken at unequal time points and (iii) missing observations. The analysis of longitudinal data should therefore take into account firstly, the within subject correlation, secondly the measurements taken at unequal time intervals and finally the missing observations. Repeated measures analysis of variance (Winer, 1971) can be used to analyse longitudinal or repeated measures data for balanced study design, i.e. when all subjects are measured at equal time points and there are no missing data. It is very rare to find balanced data sets in longitudinal studies so it is necessary to use some alternative techniques which can handle unbalanced data. Random effects or mixed effects models are one such alternative.

2.1.1 Random Effects Models

Random effects model and mixed effects model terms are usually used interchangeably. Based on the work of Harville (1977), random effects models were developed by Laird and Ware (1982). In fitting random effects models, one need to estimates parameters for fixed effects, parameters for random effects and variance-covariance parameters for within-subject and between subject variation. For balanced data it has been a common practice to estimate these parameters by equating the mean

squares in the ANOVA table to their expectations. Henderson (1953) developed the analogous techniques for unbalanced data to estimate the variance components. Henderson's (1953) described three methods of using general ANOVA technique in order to estimate variance components. All three methods of Henderson have been used extensively in a wide variety of applications. Inspite of their extensive use, all three methods suffer from some limitations of the general ANOVA method, i.e. ANOVA methods can yield negative estimators and distributional properties of the estimators are not known (Searle et al., 1992). Due to these weaknesses of ANOVA techniques, an alternative technique, based on maximum likelihood (ML) was developed for the estimation of variance component.

Maximum likelihood (ML) method was first applied to general mixed model by Hartley and Rao (1967). ML method developed by Hartley and Rao (1967) can be used for all mixed and random effects models. This technique do not suffer with the limitations of above technique given by Henderson (1953). ML estimation method has one disadvantage that it does not provide an unbiased estimate of variance in all cases (Diggle et al., 1995). The restricted maximum likelihood (REML) method was introduced by Patterson and Thompson (1971) for estimating variance components in general linear model. REML methodology provides an unbiased estimate of variance components.

There are several iterative numerical algorithms that can be used for ML or REML estimation of covariance parameters. It is hard to find a single iterative numerical algorithm for ML or REML that will be adequate for all applications. An algorithm that requires relatively few computations to converge to a ML or REML in

one situation may converge very slowly or fail in other situation (Harville, 1977). Most commonly used algorithms are EM (Expectation-Maximization) and NR (Newton-Raphson). Laird and Ware (1982) used EM-algorithm (Dempster, 1977) to fit random effects models. Lindstorm and Bates (1988) analyzed different data sets by using EM and NR algorithms, and concluded that NR algorithm is preferable to EM algorithm. Some of the advantages of using NR algorithm over EM algorithm (Lindstorm and Bates, 1988) are: (i) faster convergence of NR-algorithm; (ii) the number of iterations required for the NR algorithms is generally small compared with the number for the EM algorithm; (iii) the Hessian matrix (see Appendix A for definition) for the parameters is available at the end of the NR iterations. For these reasons the SAS (Statistical Analysis System) procedure PROC MIXED (SAS technical report - 229) uses NR algorithm.

Random effects models have been used by several authors to analyze lung function data (Laird and Ware, 1982; Kryzanowski et al., 1990; Sherrill et al., 1993; Sherrill et al., 1994; and Sherrill and Veigi, 1996). Laird and Ware (1982) utilized two-stage random effects models to analyze two longitudinal data sets. The estimates of parameters and likelihood for the random effects models were obtained via EM algorithm given by Dempster et al. (1977). The analysis of first data set was conducted to study the effects of air-pollution episodes on pulmonary functions among school children. Based on this analysis children with greatest lung function declines were identified and followed for further study and review of previous examination. The analysis of second data set was conducted to study the effect of tobacco smoke and fossil-fuel combustion products on the level and rate of development of pulmonary function.

Kryzanowski et al (1990); Sherrill et al (1993); Sherrill et al (1994); and Sherrill and Veigi (1996) analyzed lung function data from the Tuscon longitudinal study of airways obstructive disease (Lebowitz et al, 1975). In order to fit random effects models, Kalman filter approach (Jones, 1987) was used to calculate the maximum likelihood function, and the maximum likelihood estimates were obtained by using non-linear optimization program. Krzyzanowski et al. (1990) used two-stage random effects model (Laird and Ware, 1982) with first-order autoregressive error structure to analyze longitudinal data from the first nine surveys (1972-1985) of Tucson longitudinal study of airways obstructive disease. Kryzanowski et al. (1990) reported the changes in pulmonary function in subjects over 25 years of age, after the acute lower respiratory illness episodes. Their results indicted that the episodes of lower respiratory illnesses; a single episode of chest cold in men; and multiple episodes of chest cold in women are often followed by significant reduction in pulmonary function and these lower values may persist for several years. Sherrill et al (1993) demonstrated the use of random effects model (REM) to examine the association between respiratory symptoms, smoking and rate of decline of pulmonary function in elderly subjects. These elderly subjects participated in the Tucson Epidemiological Study of Airways Obstructive Disease commenced in 1972-1973 (Lebowitz et al, 1975). Sherrill et al. (1994) in another report based on the data from the same study (Lebowitz et al, 1975), examined the effects of smoking onset and cessation on FEV_1 in participants 18 yr. of age or older who reported a change in smoking habit during the study period. Sherrill et al. (1994) used random effects model for the comparison of pulmonary function data before and after changing smoking habits in the same subject, adjusting for important

covariables. When the outcome is not Gaussian, less development took place before 1986. The major difficulty with the analysis of non-Gaussian longitudinal data is the lack of a rich class of models such as multivariate Gaussian for the joint distribution of y_{it} (where y_{it} is the response for i^{th} subject at t^{th} time; $t = 1,2, \ldots, n_i$). For discrete data, it is not possible to fully determine the likelihood with first two moments. In such cases, we do need additional information about the distribution of the observations to determine the likelihood [Prentice and Zhao (1991), Fitzmaurice et al (1993), Liang et al (1992)]. In latter cases, usual procedures of maximizing the likelihood can not be used, because of too many nuisance parameters. However, likelihood methods are available for few cases of categorical data. The log-linear model for multivariate binary data is most widely used to specify a probability model for binary data in terms of canonical parameters (Bishop et al., 1975). The main limitation of the formulation given by Bishop et al. (1975) is that the interpretation of canonical parameters depends on the number of responses. In longitudinal studies, the number of observations differ across subjects so that the use of log-linear models in terms of canonical parameters is limited. Other log-linear model presentation were given by Bahdur (1961) and Fitzmauric et al. (1993). All these models suffer with the limitation that as the number of responses vary across subjects and are large, it is difficult to evaluate the likelihood. For all these reasons Liang and Zeger developed Generalized Estimating Equations (GEE) for continuous and discrete data.

2.1.2 Generalized estimating equations

In 1986, Liang and Zeger unified the regression approach for a variety of discrete and continuous responses by introducing the generalized estimating equations approach based on quasi-likelihood. The quasi-likelihood concept was introduced by Wedderburn (1974). Mean and variance of observations are needed to define the quasi-likelihood function. For independent outcome responses, quasi-likelihood by Wedderburn (1974) and the generalized linear models (GLMs) by McCullagh and Nelder (1989), unified the regression models for a variety of discrete and continuous variables. GLM made data analysis easy by providing a common set of methods regardless of the type of response. Linear, logistic, Poisson regression and some parametric survival analysis models are special cases of GLM. Liang and Zeger (1986) extended the univariate quasi-likelihood approach (McCullagh and Nelder, 1989) to incorporate correlations in the longitudinal data using multivariate quasi-likelihood methods.

Liang and Zeger (1986) re-invented the estimating equation approach for longitudinal studies and introduced generalized estimating equations (GEE), a multivariate analogue of quasi-likelihood. One advantage of GEE is that the estimated variance-covariance matrix of regression parameters is "robust". The variance-covariance matrix is "robust" because it provides a consistent estimate of the population variance-covariance matrix regardless of whether the working correlation matrix is correct or not. Liang and Zeger (1986) and Zeger and Liang (1986) presented an extension of generalized linear models to the analysis of longitudinal data when regression is the primary focus. The methods proposed by Liang and Zeger (1986)

reduce to maximum likelihood when y_{it} $(t = 1,2, \ldots, n_i)$ are Gaussian. According to notation used by Liang and Zeger (1986), let $\underline{Y}_i = (y_{i1}, \ldots, y_{in_i})'$ be the $n_i \times 1$ vector of outcome values and $\underline{X}_i = (x_{i1}, \ldots, x_{in_i})'$ be the $n_i \times p$ matrix of covariate values for the i^{th} subject $(i=1, 2, \ldots, K)$. The marginal density of y_{it} is given by :

$$f(y_{it}) = \exp [\{y_{it} \theta_{it} - a(\theta_{it}) + b(y_{it})\} \phi] \qquad \ldots (2.1)$$

where $\theta_{it} = h(\eta_{it})$ $\eta_{it} = x_{it}\beta$ and the first two moments are given by :

$$E(y_{it}) = a'(\theta_{it}), \quad var(y_{it}) = a''(\theta_{it})/\phi \qquad \ldots (2.2)$$

Liang and Zeger (1986) first presented the independence estimating equations under the working assumptions that repeated observations from a subject are independent of one another, and derived the estimators of regression parameters and the covariance matrix of regression estimates and extended the independence estimating equations to generalized estimating equations (GEE). GEE takes within-subject correlation into account which leads to estimators with higher efficiency. Let $\underline{R}(\underline{\alpha})$ be the $n \times n$ symmetric correlation matrix and $\underline{\alpha}$ be an $s \times 1$ vector which fully characterizes $\underline{R}(\underline{\alpha})$. Matrix $\underline{R}(\underline{\alpha})$ is referred as a 'working' correlation matrix. Liang and Zeger (1986) defined:

$$\underline{V}_i = \underline{A}_i^{1/2} \underline{R}(\alpha) \underline{A}_i^{1/2} / \phi \qquad \ldots (2.3)$$

where $\underline{A}_i = diag\{a''(\theta_{it})\}$. If $\underline{R}(\underline{\alpha})$ is the true correlation matrix for $\underline{Y}_i$'s, then $\underline{V}_i$ will be equal to $cov(\underline{Y}_i)$. General estimating equations given by Liang and Zeger (1986):

$$\sum_{i=1}^{K} \underline{D}_i' \underline{V}_i^{-1} \underline{S}_i = 0 \qquad \ldots (2.4)$$

where $\underline{D}_i = d\{a_i'(\theta)\}/d\beta = \underline{\Delta}_i \underline{A}_i \underline{X}_i$; for each i the $n \times n$ diagonal matrix $\underline{A}_i = \text{diag}\{a''(\theta_{ik})\}$; $\underline{\Delta}_i = diag(d\theta_{ik}/d\eta_{ik})$ is an $n \times n$ matrix and $\underline{S}_i = \underline{Y}_i - a_i'(\theta)$ is of order $n \times 1$ for the i^{th} subject. Equation *(2.4)* reduces to the independence equation if we specify $\underline{R}(\underline{\alpha})$ as the identity matrix.

Let $\hat{\underline{\alpha}}(\underline{Y}, \underline{\beta}, \phi)$ be a $K^{1/2}$-consistent (see Appendix A for definition) estimator of $\underline{\alpha}$ when $\underline{\beta}$ and ϕ are known, i.e $\hat{\alpha}$ for which $K^{1/2}(\hat{\alpha} - \alpha) = O_p(1)$ (see Appendix A for definition). Equation *(2.4)* can be expressed as a function of $\underline{\beta}$ alone:

$$\sum_{i=1}^{K} \underline{U}_i[\underline{\beta}, \hat{\underline{\alpha}}(\underline{\beta}, \hat{\underline{\phi}}(\underline{\beta}))] = 0 \qquad \ldots (2.5)$$

and $\hat{\underline{\beta}}_G$ is defined to be the solution of the above equation. Liang and Zeger (1986) showed that under mild regularity conditions as $K \to \infty$, $\hat{\underline{\beta}}_G \to \beta$ and $K^{1/2}(\hat{\beta}_G - \beta)$ is asymptotically multivariate Gaussian with zero mean and covariance matrix

$$\underline{V}_G = \lim_{K \to \infty} K \left(\sum_{i=1}^{K} \underline{D}_i' \underline{V}_i^{-1} \underline{D}_i \right)^{-1} \left\{ \sum_{i=1}^{K} \underline{D}_i' \underline{V}_i^{-1} \text{cov}(\underline{Y}_i) \underline{V}_i^{-1} \underline{D}_i \right\} \left(\sum_{i=1}^{K} \underline{D}_i' \underline{V}_i^{-1} \underline{D}_i \right)^{-1} \quad \ldots (2.6)$$

The variance estimate $\hat{\underline{V}}_G$ of $\hat{\underline{\beta}}_G$ can be obtained by replacing $cov(\underline{Y}_i)$ by $\underline{S}_i\underline{S}_i'$ and $\underline{\beta}$, ϕ, $\underline{\alpha}$ by their estimates in the expression *(2.6)*. The consistency of $\hat{\beta}_G$ and $\hat{V}_G$ depends on the correct specification of the mean, not on the correct choice of $\underline{R}$. Liang and Zeger (1986) showed with an example if the true correlation is moderate (= 0.3) then there is little difference between the generalized variance estimator of regression parameters (i.e. an estimator obtained by using GEE approach, taking into account the within subject correlation) and an estimator obtained by assuming repeated observations from a subject are independent of one another. Liang and Zeger (1986) showed when

correlation is large (≥ 0.70), then substantial improvement can be made by correctly specifying the within-subject correlation structure.

Zeger and Liang (1986) illustrated the use of GEE approach by analyzing the data from a study of the association of mother's stress and children's morbidity. Mothers were asked to keep a diary and to record daily whether their child was ill and their own relative stress. Data from the first 9 days of the diaries for 167 women with nearly complete records were used to illustrate use of the GEE method. All predictor and response variables used in the analysis were dichotomous. A logistic model was fitted for four different choices of correlations matrices: independence, i.e. repeated observations are uncorrelated; 1-dependence, i.e. correlations within responses are correlated with the most recent measurement and all other correlations are zero; stationary, i.e correlation was assumed to depend only on the time separating two observations, exchangeable, i.e. correlation between any two responses of the i^{th} individual is same; Similar results were obtained by assuming 1-dependence and stationary working correlation structures. Zeger and Liang (1986) showed that incorrectly specifying the correlation structure can give considerably incorrect results.

There are three distinct approaches to longitudinal data analysis. Zeger et al. (1988) discussed two of the approaches using the GEE to analyze the data from the Harvard Study of Air Pollution and Health and third approach was considered by Diggle et al. (1995). In first approach, Zeger et al. (1988) modelled population-averaged (PA) response as a function of covariates without explicitly accounting for heterogeneity between subjects (known as population-averaged model). In second approach, subject to subject heterogeneity was modelled explicitly by including a

random effects in the mixed model (known as subject-specific (SS) model). Zeger et al (1988) demonstrated the relationship between Population Average (PA) and Subject-Specific (SS) models using the GEE method. Data were available on 537 children, who were examined annually from age 7 to 10 years. Subjects with complete records were used for an analysis. The PA model was fitted to determine the association between mother's smoking status (dichotomous covariate: 1 if mother smoked and 0 if not), mothers age and interaction between mother's smoking status and mother's age. The SS model was fitted to determine the association between mother's smoking status, mothers age, interaction between mother's smoking status and mother's age, and random intercept which assumed to vary across subjects. The interaction was not statistically significant in both models. In PA model the mother's smoking status coefficient tells how the average risk over the population differs with change in mother's smoking status, while the same coefficient in SS model tells how one child's risk would change if his mother changes smoking status. SS model indicated that child's risk if his or her mother stopped smoking would decrease from 35% to 63% as random effect variance ranges from 0 to 4.0. PA model indicated that the rate of children's respiratory disease is approximately 35% greater for children of smoking mothers.

Diggle et al. (1995) considered the third approach to longitudinal data analysis. Diggle et al. (1995) used GEE approach to fit a transitional model to data from Indonesian children's health study (ICHS). This study was conducted in Indonesia to determine the causes and effects of Vitamin A deficiency in pre-school children. Diggle et al. assumed that the probability of respiratory infection for child i at visit j has a

direct dependence on whether or not the child had infection at visit *j-1*, as well as explanatory variable x_{ij}. One transitional model can be written as:

$$logit\ Pr(Y_{ij} = 1/ Y_{ij-1}, Y_{ij-2}, \ldots, Y_{i1}) = x_{ij}\beta + \alpha Y_{ij-1} \qquad \ldots (2.7)$$

The exponential of α (e^{α}) represents the odds of infection among children who did have the infection at the prior visit in comparison to those who did not have the infection at the prior visit.

Zeger and Liang (1992) published a review of statistical methods for the analysis of continuous and discrete longitudinal data in which different choices of models for longitudinal data were discussed.

Recently, GEE approach has been used to analyze the lung function data (Glencross et al (1997) and Christiani et al (1999)). Glencross et al (1997) studied the loss of lung function data among steel metal workers and Christiani et al (1999) studied the longitudinal decline in lung function among workers exposed to cotton dust. Glencross et al (1997) examined sheet metal workers over a 10-yr period to study the loss of pulmonary function and the loss of pulmonary function and the development of asbestosis or asbestos-related pleural fibrosis. Christiani et al (1999) analyzed longitudinal data collected at three time points over a 11-yr follow-up study of cotton textile workers in Shinghai, China. Questionnaire, spirometric testing, and endotoxins sampling were conducted at each time point. Christiani et al (1999) reported that cumulative dust was associated with 11-yr loss for FEV_1 loss after adjustment for confounders.

2.2 GOODNESS OF FIT

Assessing goodness-of-fit is an important part of any model building procedure. Goodness of fit can be defined as the degree to which a predicted value $\hat{y}_i$ agrees with an observed value y_i. To date the different procedures used to assess the goodness of fit for longitudinal models constitute, graphical methods (Diggle et al., 1988; Carter et al., 1992; Grady and Helms, 1995; Dawson et al., 1997); likelihood based procedures (Hartley and Rao, 1967); Akaike's information criterion (AIC) (Akaike, 1974); and comparison of parameter estimates in different models (Diem and Liukkonen, 1988; Manor and Kark, 1996; Jin and Sherrill, unpublished; Feng et al., 1996).

For the Gaussian-based linear mixed-effects model, both the mean and variance-covariance structure can be checked for goodness-of-fit using likelihood ratio test and Akaike's information criterion (AIC) (Akaike, 1974). In recent years Diem and Liukkonen (1988), Manor and Kark (1996), Jin and Sherrill (unpublished work), and Feng et al. (1996) compared parameter estimates in different models, while several other authors [Grady and Helms (1995) and Vonesh et al. (1996)] discussed different approaches for model selection by comparing the covariance structures.

2.2.1 Comparison of parameter estimators obtained by fitting different models:

Diem and Liukkonen (1988) compared three methods of longitudinal analysis of pulmonary function data. The first method was two-stage weighted regression method. In this method, individual least-square slopes of FEV_1 versus time for each subject was estimated at the first stage, and a regression of these slopes on various covariates was calculated by a maximum likelihood weighted regression method at the second stage. In

the second method, a random effects model was fitted with FEV_1 as a dependent variable and covariates multiplied by time. For parameter estimation, maximum likelihood with EM algorithm was used (Laird and Ware, 1981). In the third method, an attempt was made to apply an autoregressive error structure to the regression model of FEV_1 with time-weighted covariates, but this method was not successfully implemented. They had problem in estimating parameters in the third method using a SAS MACRO. During the maximization, the value of ρ, serial correlation between error terms, appeared to cycle periodically rather than to converge. They reported that the results from first two methods were similar.

Jin and Sherrill (unpublished work) compared two different methods namely "random effects model" (REM) and the "first difference method" (FDM). They compared the two methods using two computer simulated FEV_1 data sets. One of the data set did not have any missing observations, while other data set had missing observations at random. Estimates of the population slope and intercept used to generate simulated FEV_1 data sets were obtained from published prediction equation for healthy nonsmoking adult males (Sherrill et al., 1992). All variances used to simulate FEV_1 data sets were calculated from the study of healthy participants conducted in Tucson, Arizona [(Lebowitz et al. (1975) and Kundson et al. (1983)]. The two methods (REM and FDM) were compared by examining both bias and dispersion measures obtained from large simulated data sets (N=200), with different amounts of observational noise. Their study showed that the REM has the smallest mean squares error estimates for slopes at all levels of observational noise for both complete and missing data in comparison to the first difference method.

Manor and Kark (1996) compared four alternative statistical methods for the analysis of longitudinal data. Two of the methods were based on individual curve fitting while other two were based on the random effects model. They used these four methods to assess the stability of fatty acid composition of red blood cell membrane examined over a one-week period in patients hospitalized after acute myocardial infarction. First method (Method I) was based on individual curve fitting. In this method at first stage, a polynomial curve was fitted to each individual. In the second stage, the estimated coefficients from the polynomial model of stage I, were treated like response variable to study the effect of covariates. Inference on the polynomial coefficients carried out at the second stage were based on the assumption of equal variances across individuals. Method I is not suitable when observations are missing or measurements are unequally spaced. In these situations the variances of the estimated coefficients in individual curves are not equal and therefore standard univariate and multivariate methods can not be used in the second stage of Method I. So, in case of unbalanced data, the weighted analysis (Method II), was used in order to obtain more efficient estimates (Manor and Kark, 1996). Methods III and IV were based on a random effects models. In contrast to Methods I and II, Methods III and IV used the information available from all the subjects simultaneously. The parameters were estimated by the iterative techniques in Method III and IV. Manor and Kark (1996) reported that the results obtained by these four methods were different.

Feng et al (1996) compared GEE, four-stage, bootstrap with ML as the gold standard. The bootstrap technique is completely non-parametric in distribution and covariance structure specification and uses only the knowledge that clusters are the

experimental/resampling units rather than individuals. As described by Feng et al (1997), they drew a random sample of size K from the original K clusters with replacement and then fitted ordinary least squares estimates of regression parameters $\underline{\beta}$. They repeated this procedure 500 times and used the mean and covariance of the 500 $\hat{\underline{\beta}}$s as bootstrap estimate of $\underline{\beta}$ and of the covariance matrix of $\hat{\underline{\beta}}$. Bootstrap method originally developed as a robust procedure for independent observations. Kunsch (1989) extended the bootstrap method to time series data. The bootstrap technique was initially identified by Efron and Lepage (1992) as a potential tool for the analysis of correlated data. Feng et al. (1996) were the first one to apply bootstrap method to simulated correlated data with Gaussian error for balanced data. The rationale of resampling clusters for longitudinal data is similar to Kunsch's 'moving block' bootstrap method (Kunsch, 1989). Kunsch proposed for a stationary time series data to resample by blocks. A block is a group/batch of consecutive observations in time series. The idea behind resampling is that to account for dependency in the data. In time series data, the moving blocks are of fixed size, chosen by an analyst. Difficulty with resampling clustered/longitudinal data is that clusters may vary in size, i.e subjects have different number of observations. One possible way to account for the different sizes of cluster is to opt for weighted sampling or weighted bootstrap (Feng et al., 1996). Optimal weights should depend on the information provided by each cluster which depends on the covariance matrix and the cluster size. In last few years researchers have been working in the direction of extending bootstrap to longitudinal studies, but there is still work to be done in the direction of weighted bootstrap in longitudinal studies.

2.2.2 Model Selection Techniques For The Covariance Matrix For Incomplete Longitudinal Data:

Grady and Helms (1995) discussed the strategies for deciding the 'best' model and used a graphical technique for judging goodness-of-fit of covariance models for incomplete longitudinal data. Different covariance structures were applied to see which fits the data best as assessed by likelihood ratio (LR) tests, evaluation of covariance parameters and graphical methods. They provided suggestions about strategies for fitting the models for incomplete longitudinal data. They described the process of fitting the models and choosing various covariance structures with an example. They considered unstructured and structured covariance matrices for these models. Unstructured matrix is also known as the 'fully parameterized'. Structured covariance matrices are those in which one can observe a trend in the dispersion matrix and model the covariance according to the trend. Data used for model fitting were comprised of total cholesterol measures (mg/dl) from two groups of men, evaluated at nine time points, starting at month six and one year apart thereafter.

Grady and Helms (1995) concluded that (i) if the basic covariance structures provided by standard software are sufficient for the data to be analyzed, then it is not worth spending time fitting models with covariance structures other than those provided by standard software, (ii) if the interest lies in the expected value part of the model, then a basic covariance structure (i.e. unstructured, mixed model, compound symmetry and AR(1)) may be adequate, (iii) if there is an interest in the structure of the covariance matrix and the dependence of the measures over time, then it is worth while to try alternative models because they might offer information about the covariance not always available from the basic models, and (iv) graphing the covariance against lag

time offers insight into the correlation structure among repeated measures over time and aids in model selection.

Dawson et al. (1997) demonstrated how without first fitting a model, two different graphical techniques, draftman's display and parallel axis plots, can provide useful insights into the structure of the data and help us to determine an initial form of the covariance matrix. The draftsman's display is a two-dimensional array of scatter plots $X_i \times X_j$, $i = 1, 2, \ldots, p$, $j = 1, 2, \ldots, p$, $i \neq j$. In draftsman's display a p-dimensional data point $(x_1, x_2, \ldots, x_p)$ is displayed as a series of points (x_i, x_j) $i \neq j$ each plotted in the appropriate $X_i \times X_j$ co-ordinate system. Chambers et al (1983) showed that this technique is useful in detecting clustering and outliers. Parallel axis plots involve using a set of connected line segments to plot a p-dimensional data point. In this method p-parallel axes are drawn one unit apart corresponding to the variables $X_1, X_2, \ldots, X_p$. The data point $(x_1, x_2, \ldots, x_p)$ is plotted by drawing lines from the values x_i on the X_i axis to the value x_i+1 on the adjacent X_i+1, $i=1, 2, \ldots, p-1$ axis. Dawson et al (1997) chose centering and scaling technique so that properties of the covariance structure were retained in the scaled data. By plotting these scaled data using darftman's display plots and/or plots in a coordinate system with parallel axes as described above, properties of the covariance matrix can be visualized.

All the methods of goodness-of-fit reviewed above require more than one model to be fitted, e.g. likelihood based methods require nested models to be fitted; comparison of parameters estimates require more than one model to be fitted etc. The goodness-of-fit statistic was needed based solely on the model at hand and which should satisfy the properties of a reasonable goodness-of-fit statistic given by Kavlseth (1985). Based on

the definition of goodness-of-fit statistic (i.e., the degree to which a predicted value $\hat{y}_i$, agrees with an observed value y_i), Kvalseth (1985) described following properties that any reasonable goodness-of-fit statistics should possess. It should :

- have an intuitively reasonable interpretation,
- be independent of the units of measurements,
- have well defined end points corresponding to a perfect fit and a complete lack of fit, e.g. multiple correlation coefficient, R^2 $(0 \leq R^2 \leq 1)$, has value 1 corresponding to the perfect fit and 0 corresponding to complete lack of fit,
- be applicable to any type of model regardless of underlying distributional properties,
- be comparable across different models fit to the same data and
- have positive and negative residuals weighted equally.

Coefficient of determination, R^2, in multiple regression analysis, closely satisfies all of the above properties. Another measure of agreement between $\hat{y}_i$ and y_i is the concordance correlation coefficient described by Lin (1989), which is described below:

Let pairs of samples (Y_{1i}, Y_{2i}), $i=1, 2, \ldots, n$, are independently selected from a bivariate population with mean μ_1, μ_2; variances σ_1^2, σ_2^2 and correlation ρ. The estimate of the concordance correlation coefficient, which measures the closeness between bivariate readings is given by (Lin, 1989):

$$\hat{\rho}_c = 1 - \frac{n^{-1}\sum_{i=1}^{n}(Y_{1i} - Y_{2i})^2}{s_1^2 + s_2^2 + (\bar{Y}_1 - \bar{Y}_2)^2} = \frac{2\hat{\rho}s_1 s_2}{s_1^2 + s_2^2 + (\bar{Y}_1 - \bar{Y}_2)^2} \qquad \ldots(2.8)$$

where $s_k^2 = n^{-1}\sum(Y_{kl} - \bar{Y}_k)^2$ $(k=1,2)$, equation (2.8) estimates the degree to which pairs fall on a 45° line through the origin. and is the usual Pearson correlation coefficient based on sample values.

Vonesh (1992) modified it for the regression setting with $Y_{1l} = Y_l$ denoting the observed values, $Y_{2l} = \hat{Y}_l$ denoting the predicted values, and $\bar{Y}$ denoting the average of the Y_l, $\hat{Y}$ denoting the average of $\hat{Y}_l$, the estimate of ρ_c given by Vonesh (1992):

$$\hat{\rho}_c = 1 - \frac{n^{-1}\Sigma(Y_l - \hat{Y}_l)^2}{n^{-1}\Sigma(Y_l - \bar{Y})^2 + n^{-1}\Sigma(\hat{Y}_l - \hat{Y})^2 + (\bar{Y} - \hat{Y})^2} \qquad ...(2.9)$$

$$= 1 - \frac{ss(error)}{ss(Total) + \Sigma(\hat{Y}_l - \hat{Y})^2 + n(\bar{Y} - \hat{Y})^2} \qquad ...(2.10)$$

is similar to the usual R^2 value with the exception of the terms: $\Sigma(\hat{Y}_l - \hat{Y})^2$ and $n(\bar{Y} - \hat{Y})^2$ [$ss(Error) = n^{-1}\Sigma(Y_l - \bar{Y})^2$ and $ss(Total) = (n^{-1}\Sigma(Y_l - \hat{Y}_l)^2$]

Vonesh et al (1996) extended the above concordance correlation coefficient for longitudinal studies. Vonesh et al (1996) provided the concordance correlation coefficient $\hat{r}_c$ to assess the goodness-of-fit of model, another concordance correlation coefficient $r(\hat{w})$ to assess the goodness-of-fit of the covariance structure; and pseudo likelihood ratio test to test the equality between assumed (specified) and true covariance structure.

2.3 Grain Dust and Lung Dysfunction:

Farmers; grain elevators workers (working in country, terminal and transfer elevators); dock workers; feed, flour and seed-mill workers are exposed to grain dust.

Few longitudinal studies attempted to determine the predictors for annual decline in lung function among grain workers (Chan-Yeung, et al., 1981; Tabona et al., 1984; Enarson et al., 1985; Broder et al., 1985; Rosner et al, 1985, Rosner et al., 1988; Ware et al, 1989; Huy et al., 1991; McDuffie et al., 1991; Zejda et al., 1992; Pahwa et al., 1994). Chan-Yeung et al.,(1981); Tabona et al., (1984); and Enarson et al. (1985) examined the respiratory health of essentially the same group of grain handlers in the Port of Vancouver with a group of civic workers. and found that the annual decline in lung function was greater for grain workers than for the control subjects. Tabona and co-workers (1984) reported on the lung function decline of the same group of grain handlers in the Port of Vancouver over 6-yr period. In a longitudinal study over three-year period, Broder et al. (1985) studied at two occasions the respiratory symptoms and pulmonary function of grain elevator workers and control subjects. Enarson and co-workers (1985) described a dose-dependent relationship between total dust-exposure and lung function impairment.

Huy et al. (1991) studied the long-term effect of grain-dust exposure on longitudinal decline in lung function among grain elevator workers over 15-year period. They reported that there is a dose-response relationship between grain dust exposure and longitudinal change in FEV_1 and FVC and increase in respiratory symptoms. Huy et al. (1991) used multiple regression analysis technique to analyze the data, but they had good cumulative and average dust exposure estimates based on each worker's detailed job history.

McDuffie et al. (1992) reported the respiratory health status of Canadian grain elevator workers studied longitudinally at two different occasions 3-years apart. There

was significant increase in chronic respiratory symptoms of sputum and wheeze over the 3-year period (McDuffie et al., 1992). Obstructive dysfunction increased marginally over the 3-year period (McDuffie et al., 1992). It was also reported that ever smoking status and increasing years of employment in the grain industry contribute to decline in the proportion of workers with normal pulmonary function.

Zedja et al. (1992) examined the lung function among healthy grain workers from the start of employment in grain elevators. They divided the grain workers into four groups depending on the number of available tests per grain worker. Those seen at the baseline only (group I), those seen at the baseline and 1st recall (group II); those seen at the baseline, first and second recall(group III); and those seen at all the four time points (group IV). Those in group IV had better FEV_1 at the baseline compared to group II and group III. Zedja et al (1992) demonstrated a healthy workers effects in this population of new grain workers.

Some authors (Rosner et al, 1985, Rosner and Munoz, 1988, Ware et al, 1989, Pahwa et al., 1992) analyzed longitudinal pulmonary function data using first order autoregressive model. These models had previous lung function as one of the covariables and were fitted by ordinary least squares approach. These models are easy to fit but they do not consider a within-subject correlation structure. Rosner et al (1985), presented a first order autoregressive model in which linear multiple regression was used to relate change in response variables to explanatory variables. They used this method to analyze longitudinal data for the case of a continuous outcome variable (lung function) with lung function testing equally spaced over time. Authors used both time-independent and time-dependent covariates in their model. Rosner and Munoz (1988)

extended this model to include unequally spaced responses using non-linear regression methods.

Ware et al (1990) described methods for simultaneous cross-sectional and longitudinal analysis. They used the data collected on non-smoking participants in the Six Cities Study, a longitudinal study of air pollution and respiratory health conducted between 1974 and 1983. They used methods based on the first difference and first order autoregressive models to analyze the data. They provided cross-sectional and longitudinal estimates of age-related changes in the ratio FEV_1/Ht^2 (where FEV_1 is the forced expired volume in one second and Ht denotes the baseline height) using the first difference model and first order autoregressive model for men and women separately. Pahwa et al (1994) used first order autoregressive models to predict the annual loss of lung function among Canadian grain elevator workers who participated in the Grain Dust Medical Surveillance Program (Labour Canada, 1978). Pahwa et al. (1994) conducted the analysis to study the longitudinal changes in pulmonary function test values in male grain workers over a 6-yr period involving three observations and their findings were similar to Enarson et al. (1985). Pahwa et al. (1994) reported that exposure to grain dust resulted in increased decline in pulmonary function test values (FEV1 and FVC) among non-smoking, ex-smoking and current smoking grain workers. Annual loss in pulmonary function test values increased with increasing years in the grain industry and by 20 years in the industry, non-smoking and smoking grain workers had similar annual decline in lung function.

A traditional approach has been used by several authors (Glindmeyer et al 1991; Senthilselvan et al 1996; Jaakkola et al 1993; Gottlieb et al 1996) to represent the

temporal change for an individual by the least squares estimate of slope parameters obtained from linear regression of individual's measurements over time. These slope estimates were analyzed using standard statistical methods. This approach is commonly known as two-stage analysis. In the two-stage analysis suggested by Wishart (1938), mean values of linear and quadratic coefficients were used as a summary measure in the first stage instead of the slope estimates.

2.4 Longitudinal Models for Uncorrelated and Correlated Survival Data

2.4.1 Uncorrelated Survival Data

Several longitudinal studies have investigated the relationship of respiratory symptoms to annual decline in pulmonary function test variables (Chan-Yeung et al., 1981; Jakkola et al., 1993; Senthilselvan et al., 1996). Chan Yeung et al (1981) studied the lung functions and respiratory symptoms of grain workers in the port of Vancouver and civic workers in two surveys. They reported that the decline in lung function among grain workers was not correlated with the presence of respiratory symptoms when adjusted for some other covariates in the analysis of covariance. Data from two longitudinal studies, one from Cracow, Poland and other from Tuscon, Arizona in United States were analyzed by Kryzanowski et al (1990) to study the relationship between pulmonary function and changes in chronic respiratory symptoms. Kryzanowski et al (1990) reported that the number of associations between pulmonary function and symptoms are similar in both populations. Subjects with persistent symptoms had lower lung functions compared to those with remission of these symptoms. Subjects with lower initial FEV_1 values experienced onset of dyspnea more

commonly. In a longitudinal study from Montreal, Canada, new onset of dyspnea and wheeze was significantly related to the annual rate of decline in forced expiratory volume in one second (FEV_1) among non-asthmatic young adults (Jaakola et al., 1993). Results of longitudinal study of Canadian grain workers by Senthilselvan et al (1996) showed that the development of wheeze was related to lung function impairment. Jakkola et al. (1993) and Senthilselvan et al. (1996) used the similar statistical analysis called two-stage analysis technique. Sparrow et al. (1993) in a prospective longitudinal study examined the predictors of wheeze symptoms in middle-aged and older men who did not have any history of asthma or wheezing. Multiple logistic regression was used to determine the predictors of onset of wheeze symptoms. Current smoking, age, postural heart rate change at the initial examination. and higher methacholine airway responsiveness were independent predictors of the onset of the wheeze symptoms. Findings of Sparrow et al. (1993) could not be generalized to younger men and women, because there findings were based on middle-aged and older men. It has been reported that grain workers exposed to ambient concentrations of endotoxins are associated with increased prevalence of work-related respiratory symptoms (cough, phlegm, wheezing, chest tightness, and dyspnea), chronic respiratory symptoms (usual cough and phlegm production) than postal workers (Schwartz et al., 1995). Grain workers reported to have diminished measures of airway function compared with postal workers (Schwartz et al., 1995). It has been reported that chronic wheeze is prevalent among dock and grain workers (Dimich-Ward HD et al., 1995). Dimich-Ward HD et al. (1995) compared the lung functions and respiratory symptoms of dock workers who load grain cargoes with civic workers who were not occupationally exposed to grain or other dusts and with

grain workers. It was reported by Dimich-Ward HD et al. (1995) that occasional wheeze and dyspnea for the dock and grain workers were twice as great as those of civic workers. Grain dust-induced asthma has been reported by several investigators (Chan-Yeung et al., 1979; doPico et al., 1982; Warren, 1990; Chan-Yeaung, 1990). Wheeze and attacks of breathlessness are related to annual rate of decline in pulmonary function (Jedrychowski et al., 1988). Pahwa et al (1998) reported the predictors for first episode of wheezing among Canadian grain elevator workers.

2.4.2. Correlated Survival Data

In several biomedical studies, data consists of correlated failure time observations. If correlated survival times are treated as independent observations then consistent estimates of the relative risk parameters can be obtained by Cox's partial likelihood method (Wei et al, 1989; Lee et al, 1992). This is not true for the standard errors of the relative risk parameters, because the inverse of the information matrix may not be consistent estimator of the asymptotic variance. A consistent and robust estimate of the asymptotic variance was proposed by the same authors (Wei et al., 1989; Lee et al., 1992). Wei et al. (1989) was interested in testing the effectiveness of the drug ribavirin used for patients with acquired immune deficiency (AIDS) in preventing HIV-1 virus positivity over time. Blood samples for each AIDS patient were collected at weeks 4, 8, and 12. For a given blood sample, the failure time was number of days until the virus was detected in the blood sample. The data consisted of three failure times for each patient, and an example of correlated survival data. Wei et al. (1989) provided a procedure to obtain a consistent estimate of variance-covariance matrix of regression

estimates for correlated survival data. Lipsitz et al. (1994) used a fully parametric model and provided a "one-step" jackknife estimator for the variance of estimates of relative risk parameters in the analysis of correlated survival data on 271 multiple myeloma patients from 81 hospitals to compare the effects of two chemotherapy treatments on survival of the patients (Lipsitz et al, 1994). Because of the tendency for the data within hospital to be correlated, survival analysis techniques for correlated survival data were used by the authors (Lipsitz et al., 1994). In an analysis of correlated survival data, from a cross-over trial to determine the effects of placebo and different doses of a drug (ribosvrin), Lipsitz and Parzen (1996) used a semi-parametric Cox's model and obtained a completely iterative jackknife estimator for the variance of estimates of relative risk parameters. The authors also showed that the "one-step" jackknife and completely iterative jackknife estimators were asymptotically equivalent to the robust estimators proposed by Wei et al. (1989) (referred to as WLW method in this thesis) and Lee et al. (1992).

In provocation test, censoring occurs due to the failure to reach a determined endpoint i.e., usually a 20% decrease in FEV_1 at the highest dose of the agonist/stimuli. To overcome this difficulty, in recent years some researchers have used survival analysis techniques to analyse these data. Verberene et al (1993) conducted a double-blind, randomized, placebo-controlled, crossover design consisting of 17 boys and 3 girls with mild-to-moderate asthma. They studied the duration of bronchodilation and the protective effect against methacholine-induced airway obstruction of a single dose of salmetrol. Verbene et al (1993) were the first one to use survival analyses techniques to PD_{20} data. Sestini et al (1995) used Kaplan-Meier method for the comparison of

allergen-specific bronchial reactivity in different group of patients allergic to grass-pollen. They used dose of the stimulus for time and the PD_{20} for the outcome. They reported patients with rhinitis tended to reach a PD_{20} at higher doses. Sestini et al (1996) compared the results obtained using Cox proportional hazard survival analysis with the parametric tests on the slope of the dose-response curve. In patients with respiratory symptoms and a positive skin reaction to the relevant allergen, 401 allergen challenges were performed for diagnostic purposes. Based on Cox analysis, they concluded that symptoms of asthma were associated to increased risk of a reduced PD_{20} compared to rhinitis alone. They reported that no significant differences were attributable to sex, age, or to different allergens. Based on the parametric tests, their conclusion was that slope was significantly correlated to symptoms of asthma. They also reported a significant difference between rhinitis alone and asthma with or without rhinitis. They concluded that in allergen sensitivity studies, methods based on survival analysis of PD_{20} may be more sensitive than parametric tests applied to the slope of the dose-response curve.

3. DATA SETS

3.1 Introduction

In this chapter, data sets obtained from two different longitudinal studies on grain elevator workers are described. The first data set which is based on the Grain Dust Medical Surveillance Program is explained in Section 3.2. The second data set which is based on the new grain workers study is explained in Section 3.3.

3.2 Grain dust medical surveillance program

In Canada, a large number of workers are exposed to grain dust. These workers are either farmers or employees of grain elevators, feed, seed, and flour mills. The federal Government, grain dust industry employers and workers recognized the possible health hazards for workers occupationally exposed to cereal grain dust. In 1976, Labour Canada formulated the national programs for health surveillance and environmental monitoring, to obtain a complete picture of grain workers' health on a continuing basis. In 1978, Labour Canada issued the guidelines for the Environmental and Medical Surveillance Programme in the Grain Industry (File Number : 895-7-11, 1978).

3.2.1 Medical surveillance program

The Grain Dust Medical Surveillance Program (GDMSP) was a part of the above mentioned environmental and medical surveillance programs. This longitudinal

program was implemented for all employees who had been continuously employed in the grain industry for more than ninety days during a period of one year or intermittently for a total of six months in three years. Workers were tested every three years. The surveillance program was conducted in eight major phases, which are explained in chapter two of technical report of Cycle I (Labour Canada, 1984).

The medical surveillance consisted of one or more medical examinations of the worker. Medical examinations and procedures were performed by or under the direction of a licensed physician. An employee who refused to participate in the surveillance program was advised of the risk involved by his refusal. The objective of the program was to monitor the worker's health on a continuing basis while examining the association of this with the dust level of the work environment. Eight provinces and territories participated in the environmental monitoring and medical surveillance programs. These were divided into five regions: Atlantic (East of Quebec); St. Lawrence (Quebec only); Great Lakes (Ontario; East of Thunder Bay); Central (Ontario; Thunder Bay and Westward; Manitoba; and Saskatchewan); and Mountain (Alberta, British Columbia; Yukon; and North West Territories).

The Grain Dust Medical Surveillance program was commenced in 1978. Data on respiratory symptoms and pulmonary function tests were collected between 1978 and 1993 as part of the Labour Canada National Grain Dust Medical Surveillance Program (GDMSP). The data were collected in intervals called a "Cycle". The periods of cycles and the number of grain workers who participated in each cycle were: October 1978 - September 1981 - Cycle I (5702); October 1981 - September 1984 - Cycle II (5491); October 1984 - September 1987 - Cycle III (3713); October 1987 -

September 1990 - Cycle IV (2847); and October 1990 - September 1993 - Cycle V (3079).

Data were collected on the following: company, province; region; type of elevator; age; height; weight; smoking information; lung function measurements (e.g. FEV_1; FVC; MMFR; and FEV_1/FVC ratio); respiratory symptoms (e.g. chronic cough; chronic sputum; chronic wheeze; and chronic dyspnea); grade change; and physician. Chest X-rays were also available in Cycle I and Cycle II. After the completion of Cycle I, it was recommended that chest X-rays be discontinued because they were of little importance.

The data were collected in each province/territory and sent to Labour Canada. Cycle I and Cycle II data were computer coded by Labour Canada. Data for Cycles III to Cycle V were computer coded by staff of the Environmental Epidemiology Unit, Centre for Agricultural Medicine, University of Saskatchewan. The descriptive statistical analysis for Cycle I data was conducted by Labour Canada. The descriptive statistical analysis for Cycle II; III; IV and V was conducted at the Centre for Agricultural Medicine..

The questionnaire used to collect the above mentioned information and definitions of individuals terms are explained in Appendix C.

3.2.2 Environmental surveillance program

This program ran parallel to the Medical Surveillance Program. In the environmental surveillance program, dust concentrations in workplaces were monitored. All terminal, transfer and process elevators, and a representative number of country

elevators were covered by this program. The measurements suggested in the program specifications were repeated at regular intervals not exceeding two years. The specifications are given in "Guidelines for an Environmental and Medical Surveillance Program in the Grain Industry" file 897-7-11, OSHB, Labour Canada, 1978".

Labour Canada and the Grain Companies collected area and personal samples of ambient grain dust. In 1982, 68% of terminal/transfer and 69% of primary elevators were mechanically ventilated (Status Report on the Labour Canada Environmental and Medical Surveillance Program in the Grain Industry, 1982). During 1980-1984, Labour Canada staff collected and analyzed 526 grain dust samples from 17 different grain elevator companies that participated in the study. The dust samples were not randomly collected. Samples were collected by Labour Canada regional inspectional staff in response to specific complaints and at the discretion of the inspectors. A total of 340 grain dust samples from 14 terminal elevators and 186 samples from three primary elevators were collected. Only four companies (Alberta Terminals, Alberta Wheat Pool, Saskatchewan Wheat Pool and United Grain Growers) had more than 50 samples analyzed during the five year period.

3.2.3 Issues related to the Grain Dust Medical and Environmental Surveillance Programs

There were some limitations in the surveillance programs:

(i) The level of dust to which each worker was exposed at the time of spirometric tests was not available to us.

(ii) While all workers were covered by the medical surveillance program, only a sample of country elevators were covered by the environmental program.

(iii) The environmental data were available for twelve companies, while twenty five companies were included in the medical surveillance program.

(iv) Due to confidentiality reasons, no identifiers on the grain elevator workers were available to correlate with the environmental data, so these data could not be matched with individual workers' medical health information.

There were some problems associated with the collection and management of these Canada-wide health surveillance data. Some of the problems associated with the first three cycles are reported in the Programme Evaluation Report of the Environmental and Medical Surveillance Programme in the Grain Industry - Volume II - Findings (Dosman and McDuffie, 1987). While coding Cycle III data, a major systematic error was noticed with the spirometric data from Alberta. The ratio of $FEV_1/FVC*100$ was close to or equal to 100 %, which indicated technical problems with the measurements of FVC or FEV_1 or both. A detailed comparison of observed and predicted values in Saskatchewan, Alberta, and British Columbia, revealed that observed spirometric values from Alberta were significantly different from Saskatchewan and British Columbia (Dosman and McDuffie, 1987). The possible reasons for these type of errors were identified in the report by Dosman and McDuffie (1987). Because of the errors in the spirometric values, these Alberta data were deleted from the longitudinal data analysis reported by McDuffie et al. (1991) and Pahwa et al. (1994).

3.3 New grain workers' study

This was a longitudinal study on young men starting their employment histories in the grain industry. First year male college students in the College of Agriculture at the University of Saskatchewan were selected as controls. The objective of this study was to evaluate the development of respiratory symptoms and pulmonary dysfunction in new grain elevator workers.

3.3.1 Study population

All male grain workers commencing employment with the Saskatchewan Wheat Pool country elevator system in the summer and fall of 1980 were studied. A pre-employment physical examination, respiratory assessment and lung function tests were conducted on these newly recruited grain workers at the Division of Respiratory Medicine, Department of Medicine, Royal University Hospital in Saskatoon. The time-lag between pre-employment evaluation and employment commencement was approximately three months. First year College of Agriculture students at the University of Saskatchewan were chosen as control subjects. It was assumed that previous experience to grain dust would be similar in the two groups of individuals. Subjects were excluded if they had previously been grain elevator workers, had other industrial or mining exposures potentially detrimental to respiratory health, or had a respiratory infection in the previous six weeks. No subjects were excluded on the basis of antecedent chest disease.

In 1980, we studied 217 men who had just commenced work in the grain industry and 118 age-matched male control subjects (baseline). One year later (1st

recall), we re-evaluated 117 grain workers and 101 control subjects. Two years later (2nd recall), 109 grain workers and 98 control subjects were re-examined and after four years (3rd recall), 53 grain workers and 40 control subjects were re-examined. At baseline and at each recall, allergy skin prick and histamine inhalation tests were conducted on the grain elevator workers and control subjects.

3.3.2 Respiratory symptoms' questionnaire

With the assistance of a trained technician, both the grain workers and control subjects completed a questionnaire based on the American Thoracic Society (ATS) questionnaire (Ferris, 1978). In the first part of the questionnaire, demographic information e.g age, height, weight, name was collected. A unique identification number was given to each subject. All the information collected was kept confidential. Detailed information was collected on respiratory symptoms e.g. cough; wheeze; sputum; shortness of breath; and whether exposure to grain dust caused any illnesses e.g. skin; nasal; or eye irritation. Questions were also asked about their past medical history and whether any first or second degree relatives had any history of respiratory diseases or asthma. An allergy skin prick test (Section 3.3.3) and a histamine challenge test (Section 3.3.4) were done on each grain worker. The study protocol, questionnaire and correspondence were approved by the University of Saskatchewan's President's Advisory Committee on Ethics in Human Experimentation.

The questionnaire used in this study is attached in Appendix C.

3.3.3 Allergy skin prick test

Allergy skin-prick tests were conducted on each worker and control subject with a battery of 14 allergens (Hollister-Stier Division of Miles Labs Ltd, Rexdale, Ontraio) and a dilutent control. The following allergens were used for the skin test: Alternaria, Aspergillus, Penicillium, Hormodendrum, house dust mite, mixed animal dander, mixed grass pollen, mixed weed pollen, mixed tree pollen, wheat dust extract, rye dust extract, oat dust extract, barley dust extract, and rapeseed extract. The wheal diameters were measured in two perpendicular directions at 10 minutes and the mean wheal diameter was determined as the mean of the measurements. A positive skin-prick test was recorded when the mean of the measurements was equal to or greater than 2 mm. A form used to record allergy skin prick test results is attached in Appendix C.

3.3.4 Assessment of bronchial hyperresponsiveness

Bronchial responsiveness was evaluated by the method of Cockroft et al (1977). The subjects began by inhaling normal saline. The FEV_1 measured following inhalation of normal saline was used as a "control". The subject inhaled sequentially increased doses of histamine using the following dose increments: 0.00625 mg/ml; 0.0125 mg/ml; 0.25 mg/ml; 0.5 mg/ml; 1.0 mg/ml; 2.0 mg/ml; 4.0 mg/ml; 8.0 mg/ml. A positive test resulted when inhalation of histamine produced a 20% decrease in the FEV_1. The values for PC_{20} were either interpolated from a log concentration-response curve or form the last two responses, when appropriate. A 20% fall in FEV_1 is the most commonly used clinical response. The histamine concentration at which the 20% decrease occurs is

referred to as the provocation dose/concentration (PD_{20}/PC_{20}), which is used as an indicator of bronchial hyperresponsiveness.

3.3.5 Strengths of the New Grain Workers' Study

This study had two major strengths. The first strength of this study was that the same equipment was used on all the occasions. Another strength was that all the lung function measurements, allergy skin test, and bronchial hyperresponsiveness tests after baseline were conducted by one technician. At the baseline, these tests were administered by three technicians. Data management and documentation was done by one individual with appropriate quality control measures.

PART I - OUTCOME - CONTINUOUS VARIABLE

4. METHODS - MODELS FOR LONGITUDINAL DATA ANALYSIS

4.1 Introduction

We are often interested in examining the relationship between a response (Y) and a set of explanatory variables (X). When the responses are independent and Gaussian, the approach of utilizing classical/general linear models is well developed based on the mathematical theory of normal distribution. When the responses are independent, a class of regression models for Gaussian and non-Gaussian data, known as generalized linear models, based on quasi-likelihood extends the scope of examining the relationship between Y's and X's. The generalized linear models (GLMs) for independent responses are an extension of the classical/general linear models for independent data. The standard GLM analyses techniques cannot be used to analyse longitudinal data, because in longitudinal studies, repeated observations on a subject tend to be correlated and the assumption of independence is invalid in longitudinal studies. In the analysis of longitudinal data, the inter-dependence among repeated observations must be taken into account to make valid inferences about the relationship between Y with X. Extensions of classical/general linear models for longitudinal data are well developed using the mathematical theory of multivariate normal distribution. Extensions of GLMs based on multivariate quasi-likelihood known as Generalized Estimating Equations allow us to model continuous and discrete responses. Different

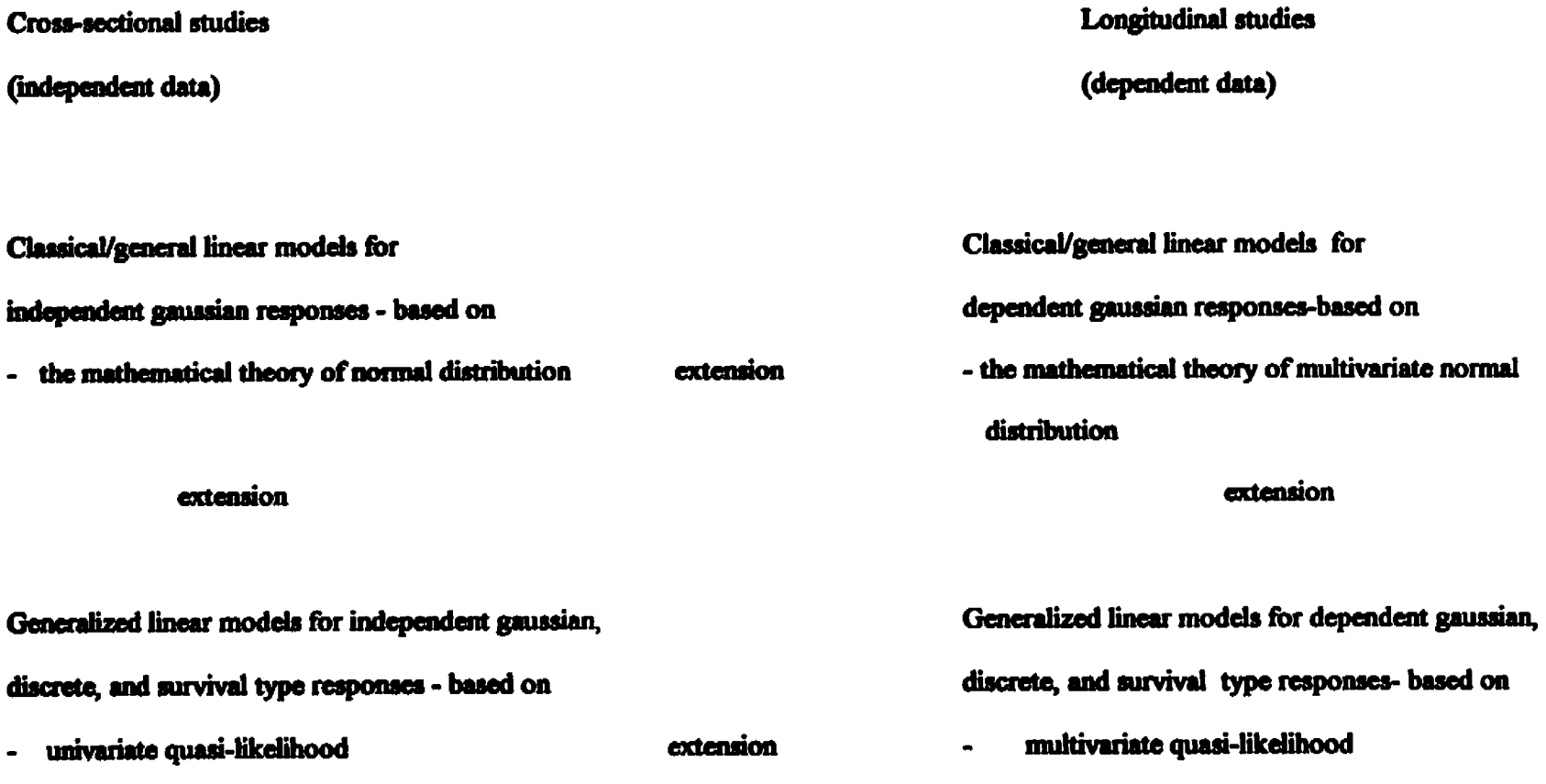


Fig. 1. Statistical methods used for independent and dependent responses

statistical models used for dependent and independent data are summarized in Figure 4.1.

Some of the mathematical theory of general linear models for gaussian longitudinal data given by Diggle et al. (1995). Diggle et al. (1995) discussed the different approaches to parameter estimation based on the multivariate normal distribution. Diggle et al. (1995) formulated the weighted least squares for the estimation of regression parameters; maximum likelihood and restricted maximum likelihood approaches for the estimation of covariance parameters. The definition of univariate quasi-likelihood equations for parameter estimation of independent observations are given in Section 4.2. The generalized linear model for longitudinal data is described in Section 4.3. Assessing goodness-of-fit is an important part in any model building. For gaussian longitudinal data most commonly used goodness-of-fit statistics are the likelihood ratio test and Akaike information criterion which are given in Sections 4.8.1 and 4.8.2 respectively. Vonesh et al. (1996) developed concordance coefficients to assess the goodness-of-fit for longitudinal models and variance-covariance structures. They also developed a pseudo-likelihood ratio test for testing the null hypothesis that the assumed covariance structure is equal to the true covariance structure. This approach of assessing goodness-of-fit for longitudinal models is still in its early stages. These goodness-of-fit statistics are described in Sections 4.8.3 and 4.8.4.

4.2. Univariate quasi-likelihood function

Suppose we have independent observations y_i ($i = 1,2, \ldots, n$) with expectations μ_i and variance $V(\mu_i)$, where V is some known function. We also suppose that for each observation, μ_i is some known function of a set of parameters $\beta_1, \beta_2, \ldots, \beta_p$. Then for each observation, the quasi-likelihood function $Q(y_i, \mu_i)$ (Wedderburn, 1974) is defined by the relation:

$$U = \frac{\partial Q(y_i, \mu_i)}{\partial \mu_i} = \frac{y_i - \mu_i}{V(\mu_i)} \qquad \ldots (4.1)$$

$$Q(y_i, \mu_i) = \int^{\mu} \frac{y_i - \mu_i}{V(\mu_i)} d\mu_i \qquad \ldots (4.2)$$

Q has many properties similar to a log likelihood function. The common properties of U or Q were given by McCullagh and Nelder (1989) and Wedderburn (1974).

The quasi-likelihood estimating equations for the regression parameters $\underline{\beta}$, obtained by differentiating $Q(y; \mu)$, may be written in the form $U(\underline{\beta}) = \underline{0}$, where

$$U(\underline{\beta}) = \sum_{i=1}^{n} (\partial \mu_i / \partial \underline{\beta}) (Var(Y_i))^{-1} (Y_i - \mu_i) = \underline{0} \qquad \ldots (4.3)$$

is called the quasi-score function.

4.3. Generalized Linear Models for Dependent Observations

Three extensions of generalized linear models for longitudinal data are marginal, transitional, and random effects models. A multivariate analogue of quasi-likelihood is generalized estimating equations (GEE). The continuous and discrete longitudinal data can be analyzed by using the GEE approach. Some of the matrices and vectors commonly used in the theory of longitudinal data analysis are given in Section 4.3.1.

4.3.1. Notations of matrices and vectors

Let there be m subjects and n_i observations on i^{th} subject in a longitudinal study, i.e. $i = 1, 2, ..., m$ subjects; $j = 1, 2, ..., n_i$ responses for i^{th} subject recorded at times $t_{i1} < t_{i2} < ... < t_{in_i}$.

Let Y_{ij} be the observed response for subject i at time t_{ij} ($i = 1, 2, ..., m; j = 1, 2, ..., n_i$). So,

$\begin{bmatrix} y_{i1} \\ y_{i2} \\ \vdots \\ y_{in_i} \end{bmatrix}$ is the $n_i \times 1$ column vector containing the n_i responses for subject i.

$\begin{bmatrix} y_{11} \\ \vdots \\ y_{1n_1} \\ y_{21} \\ \vdots \\ y_{2n_2} \\ \vdots \\ y_{m1} \\ \vdots \\ y_{mn_m} \end{bmatrix}$ is the $(\Sigma n_i) \times 1$ column vector containing the responses for all the subjects.

Let X_{ij1}, X_{ij2}, *and* X_{ijp} be the values taken by the p covariates $X_1, X_2, ..., X_p$ for the i^{th} subject at time t_{ij}. So,

$\underline{X}_{ij} = \begin{bmatrix} X_{ij1} \\ X_{ij2} \\ \vdots \\ X_{ijp} \end{bmatrix}$ is the $p \times 1$ column vector of covariate values for the i^{th} subject at time t_{ij}

and

$$\underline{X}_i' = \begin{bmatrix} \underline{X}_{i1}' \\ \underline{X}_{i2}' \\ \cdot \\ \cdot \\ \cdot \\ \underline{X}_{in_i}' \end{bmatrix} = \begin{bmatrix} X_{i11} X_{i12} \ldots X_{i1p} \\ X_{i21} X_{i22} \ldots X_{i2p} \\ \cdot \\ \cdot \\ \cdot \\ X_{in_i 1} X_{in_i 2} \ldots X_{in_i p} \end{bmatrix}$$ is the $n_i \times p$ matrix of covariate values for the i^{th} subject ; and

$$\underline{\varepsilon}_i = \begin{bmatrix} \varepsilon_{i1} \\ \varepsilon_{i2} \\ \cdot \\ \cdot \\ \cdot \\ \varepsilon_{in_i} \end{bmatrix}$$ is $n_i \times 1$ vector of random errors for the i^{th} subject..

In longitudinal data analyses, it is important to consider all possible correlations among the n_i responses: $\{ Y_{i1}, Y_{i2}, \ldots, Y_{in_i} \}$ for the i^{th} subject. There are $n_i(n_i-1)/2$ such correlations, one for each distinct pair of responses. So,

$$\underline{R}_i = \begin{bmatrix} 1 & corr(Y_{i1},Y_{i2}) & \ldots & corr(Y_{i1},Y_{in_i}) \\ corr(Y_{i2},Y_{i1}) & 1 & \ldots & corr(Y_{i2},Y_{in_i}) \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ corr(Y_{in_i},Y_{i1}) & corr(Y_{in_i},Y_{i2}) & \ldots & 1 \end{bmatrix}$$

is an $n_i \times n_i$ matrix of within-subject correlations. The entry in the j^{th} row and k^{th} column of $\underline{R}_i$ is $corr(Y_{ij}, Y_{ik})$. In any longitudinal analysis it is necessary to consider the appropriate structure of $\underline{R}_i$ for each subject, or equivalently, the structure of variance-covariance matrix $\underline{\Sigma}_i$ of the responses $\{Y_{i1}, Y_{i2}, \ldots, Y_{in_i}\}$ for the i^{th} subject, $i = 1, 2, \ldots, m$. As the $corr(Y_{ij}, Y_{ik}) = cov(Y_{ij}, Y_{ik}) / \sqrt{Var(Y_{ij})\, Var(Y_{ik})}$, so we can write $\underline{\Sigma}_i = \underline{V}_i^{1/2} \underline{R}_i \underline{V}_i^{1/2}$, where

$$\underline{V}_i^{1/2} = \begin{bmatrix} \sqrt{Var(Y_{i1})} & 0 & 0 \\ 0 & \sqrt{Var(Y_{i2})} & 0 \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ 0 & 0 & \sqrt{Var(Y_{in_i})} \end{bmatrix}$$

$= Diag\,[\sqrt{Var(Y_{i1})}, \sqrt{Var(Y_{i2})}, \ldots, \sqrt{Var(Y_{in_i})}]$, so

$$\underline{\Sigma}_i = V(Y_i) = \begin{bmatrix} var(Y_{i1}) & cov(Y_{i1},Y_{i2}) & \cdots & cov(Y_{i1},Y_{in_i}) \\ cov(Y_{i1},Y_{i2}) & var(Y_{i1}) & \cdots & cov(Y_{i2},Y_{in_i}) \\ \vdots & \vdots & \ddots & \\ cov(Y_{i1},Y_{in_i}) & cov(Y_{i2},Y_{in_i}) & \cdots & var(Y_{i1}) \end{bmatrix}$$

This equation relates $\underline{\Sigma}_i$ and $\underline{R}_i$.

While modelling longitudinal data, the primary objective of regression analysis is to identify the relationship between the expected value E(Y) of the response variable Y and the covariates $X_1, X_2, \ldots, X_p$. Modelling the correlation structure is of secondary importance, however it is necessary to take into account any intra-subject response correlation when making statistical inferences about the regression coefficient $\beta_1, \beta_2, \ldots, \beta_p$. If we do not take into account the intra-subject correlation, then such statistical inferences can be seriously in error. Some of the most commonly used within-subject correlation matrices are as follows:

1) independence, i.e. repeated observations are uncorrelated.

$$\underline{R} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}$$

2) unspecified (unstructured), i.e. correlations within any two responses are unknown and need to be estimated.

$$\underline{R} = \begin{bmatrix} 1 & \rho_{1,2} & \cdots & \rho_{1,n} \\ \rho_{1,2} & 1 & \cdots & \rho_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{1,n} & \rho_{2,n} & \cdots & 1 \end{bmatrix}$$

3) exchangeable, i.e. correlation between any two responses of the i^{th} individual is the same.

$$\underline{R} = \begin{bmatrix} 1 & \rho & \rho & \cdots & \rho \\ \rho & 1 & \rho & \cdots & \rho \\ \rho & \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \rho & \cdots & 1 \end{bmatrix}$$

4) autoregression of first order [AR(1)] assuming the sampling interval length is the same between any two observations

$$\underline{R} = \begin{bmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{n-1} \\ \rho & 1 & \rho & \cdots & \rho^{n-2} \\ \rho^2 & \rho & 1 & \cdots & \rho^{n-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{n-1} & \rho^{n-2} & \rho^{n-3} & \cdots & 1 \end{bmatrix}$$

5) Autoregression of the first order assuming continuous unequally spaced sampling intervals $\{t_1, t_2, \ldots, t_n\}$

$$\underline{R} = \begin{bmatrix} 1 & \rho^{|t_2-t_1|} & \rho^{|t_3-t_1|} & \cdots & \rho^{|t_n-t_1|} \\ \rho^{|t_2-t_1|} & 1 & \rho^{|t_3-t_2|} & \cdots & \rho^{|t_n-t_2|} \\ \rho^{|t_3-t_1|} & \rho^{|t_3-t_2|} & 1 & \cdots & \rho^{|t_n-t_3|} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{|t_n-t_1|} & \rho^{|t_n-t_2|} & \rho^{|t_n-t_3|} & \cdots & 1 \end{bmatrix}$$

4.3.2. Multivariate quasi-likelihood function

The multivariate quasi-likelihood function (Liang and Zeger, 1986) is an extension of the univariate quasi-likelihood function (Wedderburn, 1974). The following notations will be used to define the multivariate quasi-likelihood function:

- Response vector:

  $\underline{y}_i = (y_{i1}, y_{i2}, \ldots, y_{in_i})'$, $i = 1, 2, \ldots, m$

- Mean Vector

  $E(\underline{y}_i) = \underline{\mu}_i = (\mu_{i1}, \mu_{i2}, \ldots, \mu_{in_i})'$

- Predictors:

  $X_{ij} = (X_{ij1}, X_{ij2}, \ldots, X_{ijp})'$, $j = 1,2 \ldots, n_i$; $i = 1, 2, \ldots, m$

- Link function:

  $g(\mu_{ij}) = X_{ij}' \underline{\beta} = \Sigma X_{ijk} \beta_k$, $j = 1, 2, \ldots, n_i$; $i = 1, 2, \ldots, m$

- Variance: Let $\underline{\Sigma}_i$ be the $n_i \times n_i$ variance-covariance matrix for $\underline{Y}_i$, and $\underline{R}_i$ be the $n_i \times n_i$ corresponding correlation matrix for $\underline{Y}_i$ and $\underline{V}_i^{1/2}$ is a diagonal matrix whose j^{th} diagonal element is given by the expression $\sqrt{(\phi V(\mu_{ij}))}$, so

  $\text{Var}(Y_{ij}) = \phi V(\mu_{ij})$ and

  $\text{Var}(\underline{Y}_i) = \underline{\Sigma}_i = \underline{V}_i^{1/2} \underline{R}_i \underline{V}_i^{1/2}$

Using the above notations, the multivariate quasi-likelihood function is defined as:

$$\underline{U}(\underline{\beta}) = \sum_{i=1}^{m} \left[\frac{\partial \underline{\mu}_i}{\partial \underline{\beta}}\right] [\underline{\Sigma}(\underline{\alpha})]^{-1} [\underline{Y}_i - \underline{\mu}_i] \quad \ldots (4.4)$$

where,

$\partial \underline{\mu}_i / \partial \underline{\beta}$ is $p \times n_i$ matrix and $\partial \underline{\mu}_i / \partial \underline{\beta} = ((\partial \mu_{ij} / \partial \beta_k))$, $i = 1,2, \ldots, m$; $j = 1,2, \ldots, n_i$; $k = 1,2, \ldots, p$

$\underline{\Sigma}(\alpha)$ is $n_i \times n_i$ matrix and $\underline{\Sigma}(\alpha) = \underline{V}_i^{1/2} \underline{R}_i \underline{V}_i^{1/2}$

$(\underline{Y}_i - \underline{\mu}_i)$ is $n_i \times 1$ vector and $(\underline{Y}_i - \underline{\mu}_i) = (Y_{i1} - \mu_{i1}, Y_{i2} - \mu_{i2}, \ldots, Y_{in} - \mu_{in})$

The common properties of the multivariate quasi-likelihood function, $\underline{U}(\beta)$, were described by McCullagh and Nelder (1989).

4.3.3. Multivariate quasi-likelihood equations/generalized estimating equations

Using the above notations, multivariate quasi-likelihood estimating equations are given by:

$$\sum_{i=1}^{m} \left[\frac{\partial \mu_i}{\partial \beta}\right] [\underline{\Sigma}(\alpha)]^{-1} [\underline{Y}_i - \underline{\mu}_i] = \underline{0} \qquad \ldots (4.5)$$

Equations given in *(4.5)* are also known as Generalized Estimating Equations.

4.4 Marginal, transitional, and random effects models for longitudinal data

Marginal, transitional and random effects models are three extensions of the generalized linear model (GLM). In practice, there are usually four types of responses: continuos, discrete, count, or survival type. In this thesis, models are fitted for continuous (FEV_1 and FVC) response outcomes; discrete (bronchial hyperesponsives: present or absent), and survival type (bronchial hyperresponsiveness: dose for 20% fall in FEV_1) outcomes. We fit linear models for continuous longitudinal data and non-linear models for discrete longitudinal data (e.g. logistic model for binary data). In linear models, $E(Y_{ij})$ is a linear function of the regression coefficients $\beta_1, \beta_2, \ldots, \beta_{p-1}, \beta_p$. For discrete data, $E(Y_{ij})$ is not a linear function of $\beta_1, \beta_2, \ldots, \beta_{p-1}, \beta_p$. It is much more difficult to fit some types of models using discrete data, because variance is a

function of the mean, e.g.; for a binomial distribution: E(Y) = np and Var(Y) = $np(1-p) = np(1 - np/n) = E(Y)[1 - E(Y)/n]$

The marginal, transitional and random effects models for continuous outcomes, and the marginal model for discrete outcomes are explained below. The generalized estimating equations approach was used to analyze data obtained from the Grain Dust Medical Surveillance Program (GDMSP) (described in Section 3.1, Chapter 3), by fitting marginal, and transitional models assuming different covariance structures. Random effects models were fitted by using PROC MIXED which uses the maximum likelihood technique or restricted maximum likelihood for variance components estimation and Henderson's ANOVA technique for weighted/generalized estimation of regression parameters (SAS Technical Report P-229, 1992).

4.5. Marginal models

When a population is of primary interest, fitting marginal models is the most appropriate. In these models, the population-averaged response is modelled as a function of the covariates. The regression coefficients are interpreted for the population rather than for individuals, so these are known as "population-averaged" (PA) models. In marginal models, the regression and within subject correlation are modelled separately. In this approach we assume:

M-1. Let Y_{ij} be the response for i^{th} subject at j^{th} time, then the marginal expectation $E(Y_{ij}) = \mu_{ij}$ is related to covariates x_{ij} by

$$g(\mu_{ij}) = x_{ij}'\beta \qquad \ldots (4.6)$$

where g is a known link function, such as an identity function for Gaussian responses and logit function for binary responses and log for counts.

M-2. The marginal variance is a function of the marginal mean, i.e.,

$$Var(Y_{ij}) = \upsilon(\mu_{ij})\,\phi \qquad \ldots (4.7)$$

where υ is a known function and ϕ is the over-dispersion parameter which accounts for the variation of Y_{ij} not explained by $\upsilon(\mu_{ij})$ and need to be estimated.

M-3. The covariance between Y_{ij} and the Y_{ik}, $j<k$ is a function of the marginal means and additional parameters, α, i.e.

$Cov(Y_{ij}, Y_{ik}) = \rho(\mu_{ij}, \mu_{ik}, \alpha)$, where ρ is a known function ... (4.8)

4.5.1. Marginal models for continuous outcome variable

Let $\underline{Y}_i$ be the vector of n_i responses of i^{th} subject. Y_{ij} denotes the j^{th} response of the i^{th} subject. We model the marginal expectation, $E(Y_i)$ as a function of explanatory variables, i.e. for i^{th} subject, we predict mean response at time t_{ij} as a function of covariates.

$$E(\underline{Y}_i) = \underline{X}_i\underline{\beta} \text{ or } \underline{Y}_i = \underline{X}_i\underline{\beta} + \underline{\varepsilon}_i \quad E(\underline{\varepsilon}_i) = 0 \qquad \ldots (4.9)$$

Where $\underline{Y}_i$ is $n_i \times 1$ the response vector of n_i observations for i^{th} subject

$\underline{X}_i$ is $n_i \times p$ matrix of p covariates for i^{th} subject

$\underline{\beta}$ is a $p \times 1$ vector of unknown parameters

$\underline{\varepsilon}_i$ is a $p \times 1$ vector of random errors

$$\begin{bmatrix} Y_{i1} \\ Y_{i2} \\ \vdots \\ Y_{in_i} \end{bmatrix} = \begin{bmatrix} x_{i11} & x_{i12} & \cdots & x_{i1p} \\ x_{i21} & x_{i22} & \cdots & x_{i2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{in_i1} & x_{in_i2} & \cdots & x_{in_ip} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_2 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_p \end{bmatrix}$$

The above equation is that of a classical multiple linear regression model for the i^{th} subject. In contrast, the responses $[Y_{i1}, Y_{i2}, \ldots, Y_{in_i}]$ for the i^{th} subject cannot be assumed to be mutually independent. So, in addition to modelling the marginal relationship between the Y_{ij} and the $\underline{X}_{ij}$, we should model the dependencies among the Y_{ij} for subject i. We can model these dependencies by specifying a variance-covariance matrix $\underline{\Sigma}_i$ for $\underline{Y}_i$, or a correlation matrix $\underline{R}_i$ for $\{Y_{i1}, Y_{i2}, \ldots, Y_{in_i}\}$. In this approach, the regression and within-subject correlation are modelled separately.

4.5.2. Marginal models for discrete outcomes

As previously described for marginal linear models, the goal here is to model the marginal mean or expectation $E(Y_{ij})$ as a function of important covariates. Most commonly used models for discrete outcomes are logistic models for dichotomous and polytomous outcomes, and Poisson regression models for counts.

Consider a dichotomous response variable Y_{ij}, which takes the value 1 if the i^{th} subject has a disease and 0 otherwise. The marginal mean or expectation is modelled as a function of covariates. Suppose there are two covariates, x_{1ij} and x_{2ij} for the i^{th} subject at the j^{th} time point. Again, let us assume that x_{1ij} is a dichotomous covariate (i.e $x_{1ij} = 1$ or 0 depending on whether the i^{th} subject has a particular characteristic at the j^{th} time point or not) and x_{2ij} is continuous covariate. Let $\mu_{ij} = E(Y_{ij}) = Pr\ (Y_{ij} = 1)$, then the marginal model for the dichotomous outcome is a logistic model, which can be written as:

$$logit\ \mu_{ij} = logit\ [E(Y_{ij})] = logit\ [pr(Y_{ij}=1)] = \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij} \quad \ldots (4.10)$$

In this model, we assume:

$Var(Y_{ij}) = \mu_{ij}(1 - \mu_{ij})$

$Corr(Y_{ij}, Y_{ik}) = \alpha_{|j-k|}$, $j < k$, i.e. within-subject correlation is a function of time between two repeated observations on the same subject.

For the marginal model *(4.10)*, regression coefficients have population-averaged interpretation. The parameter β_1 describes the effect of x_{1ij} on the marginal expectation of the *Y's*. The prevalence odds ratio e^{β_1}, which is the ratio of the odds of disease among subjects who have a particular characteristic compared to the odds of disease among subjects who do not have a characteristic and can be expressed as:

$$e^{\beta_1} = \frac{pr(Y_{ij}=1|x_{1ij}=1)/pr(Y_{ij}=0|x_{1ij}=1)}{pr(Y_{ij}=1|x_{1ij}=0)/pr(Y_{ij}=0|x_{1ij}=0)} \qquad \ldots (4.11)$$

β_2 describes the effect of continuous a covariable on the marginal expectation of the Y's. The magnitude of within-subject correlation i.e $corr(Y_{ij}, Y_{ik}) = \alpha$ does not alter the interpretation of β_1 and β_2.

The details of the generalized estimating equations approach for marginal model are given by Zeger and Liang (1992) and Diggle et al. (1995).

4.6. Transitional models

When the time dependence is central, models for the conditional distribution of Y_{ij} given $Y_{ij-1}, Y_{ij-2}, \ldots, Y_{i1}$ may be more appropriate. These are also known as conditional models. Transitional models are different from marginal models in that they attempt to model both the regression coefficient and the within-subject correlation simultaneously.

Assumptions of the transitional model:

1. The conditional expectation of Y_{ij}:

$E(Y_{it}/Y_{it-1}, Y_{it-2}, \ldots, Y_{i1}) = \mu_{it}^{c}$ depends on x_{it} and the past history of responses, which can be represented as:

$$g(\mu_{it}^{c}) = x'_{it}\beta + \Sigma \alpha_j f_j(Y_{it-1}, \ldots, Y_{i1}), \quad \ldots (4.12)$$

where $\{f_j, j = 1, \ldots, v\}$ are known functions.

2. The conditional variance of Y_{it} given the past history is a function of μ_{it}^{c}; that is

$$\text{var}(Y_{it} \mid Y_{it-1}, \ldots, Y_{i1}) = v(\mu_{it}^{c})\,\phi \qquad \ldots (4.13)$$

where v is a known function, and ϕ is the over-dispersion parameter.

4.6.1. Transitional models for continuous outcome variables

Linear models for the conditional mean Y_{ij} given the observed value $Y_{i,j-1}$ of the response immediately preceding Y_{ij}, i.e.

$$E(Y_{ij}/Y_{i,j-1}) = X_{ij}\beta + \rho\,(y_{i,j-1} - X_{i,j-1}\,\beta) \qquad \ldots (4.14)$$

These type of models are also called first-order autoregressive models. In general, a transitional model specifies a generalized linear model for the conditional distribution of Y_{ij} given the past responses $H_{ij} = \{Y_{ik}, k=1, \ldots, j-1\}$, and conditional expectation of $Y_{ij} = E(Y_{ij}/ Y_{ij-1}, \ldots, Y_{i1}) = \mu_{ij}$ depends on x_{ij} and past responses. It can be expressed as:

$$E(Y_{ij}/Y_{i,j-1}) = X_{ij}\beta + \rho_1\,(y_{i,j-1} - X_{i,j-1}\,\beta) + \ldots + \rho_{j-1}\,(y_{i,j-1} - X_{i,j-1}\,\beta) \qquad \ldots (4.15)$$

Generalized estimating equations approach for transitional models is given by Diggle et al. (1995).

4.7. Random effects models

Random-effects/mixed-effects models are more appropriate for the study of an individual's growth. These models are known as "subject-specific" (SS) models. Similar to the transitional models, random effects models (REM) model regression coefficients and the within-subject correlation simultaneously.

4.7.1 Random effects models for continuous outcome variables

A general form of the random-effects model (REM) for longitudinal data analysis formulated by Harville (1977) is of the following form:

$$\underline{y}_i = \underline{X}_i\underline{\beta} + \underline{Z}_i\underline{\gamma}_i + \underline{\epsilon}_i \quad \text{for } i^{th} \text{ subject} \qquad \ldots (4.16)$$

and $$E(\underline{y}_i) = \underline{X}_i\underline{\beta} \text{ and } \underline{\Sigma}_i = V(\underline{y}_i)$$

Where $\underline{y}_i$ is an $n_i \times 1$ column vector of responses for the i^{th} subject.

$\underline{\beta}$ is a p×1 vector of fixed unknown coefficients.

$\underline{\gamma}_i$ is a q×1 unknown vector of random effects (random across subjects)

$\underline{X}_i$ is $n_i \times p$ known design matrix

$\underline{Z}_i$ is $n_i \times q$ known design matrix

Let, $\underline{\gamma}_i \sim MVN(\underline{0}, \sigma^2\underline{B})$, $\underline{\epsilon}_i \sim MVN(\underline{0}, \sigma^2\underline{W}_i)$ and

$V(\underline{y}_i) = \underline{\Sigma}_i = \underline{Z}_i V(\underline{\gamma}_i) \underline{Z}_i + V(\underline{\epsilon}_i) = (\underline{Z}_i \underline{B} \underline{Z}_i + \underline{W}_i) \sigma^2$, assuming $\underline{\gamma}_i$ and $\underline{\epsilon}_i$ are independent. So, $\sigma^2\underline{B}$ is a between subject covariance matrix, $\sigma^2\underline{W}_i$ is a within subject covariance matrix. $\underline{B}$ and $\underline{W}_i$ are non-singular. The model is computationally feasible, by assuming $\underline{B}$ is small enough to be inverted easily. If $\underline{W}_i = \sigma^2\underline{I}$, and the random effect term is excluded, the model *(4.16)* reduces to the classical general linear model. Laird

and Ware (1982) assumed that $\underline{W}_i$ is parameterized in terms of a vector of unknown parameters $\underline{\theta}$.

Our aim is to estimate i) a vector of fixed effects, $\underline{\beta}$; ii) a parameter vector of random effects $\underline{\gamma}$, and iii) the variance and covariance parameters of $\underline{B}$ and $\underline{W}_i$. The fixed effects vector $\underline{\beta}$, was estimated using generalised least squares; the random effects vector $\underline{\gamma}$ was estimated using an extended version of the Gauss-Markov theorem for random effects, and the variance components were obtained by using restricted maximum likelihood, via the Newton-Raphson algorithm. SAS PROC MIXED was used to fit the random effects model.

4.7.1.1 Estimation of $\underline{W}$ and $\underline{B}$ in the mixed model:

In the general linear model, we have to estimate unknown parameters in $\underline{\beta}$. Estimation is more difficult in the mixed model, because in addition to $\underline{\beta}$, unknown parameters $\underline{\gamma}$, $\underline{W}$ and $\underline{B}$ have to be estimated. Least squares is not the suitable estimation method for mixed models. Generalized least squares (GLS) is more appropriate minimizing:

$$(\underline{y} - \underline{X}\underline{\beta})' \underline{\Sigma}^{-1} (\underline{y} - \underline{X}\underline{\beta}) \qquad ...(4.17)$$

However, the above equation requires the knowledge of $\underline{\Sigma}$ and therefore, knowledge of $\underline{W}$ and $\underline{B}$. If this information is not available, one approach is to use estimated GLS, in which some reasonable estimate of $\underline{\Sigma}$ is inserted in order to minimize the above equation. Then the goal is to find a reasonable estimate of $\underline{W}$ and $\underline{B}$. So, the parameters of $\underline{\Sigma}$ are estimated by using the normal theory maximum likelihood (ML) or restricted maximum likelihood (REML). Our aim is to optimize either the maximum likelihood

(ML) or restricted maximum likelihood (REML) log likelihood functions. In SAS procedure PROC MIXED, an objective function associated with ML or REML is constructed and maximized over all unknown parameters. The corresponding ML and REML log-likelihood functions are as follows:

$$ML: l(\underline{W}, \underline{B}) = -\frac{1}{2}\log|\underline{\Sigma}| - \frac{1}{2}\underline{r}' \underline{\Sigma}^{-1} \underline{r} - \frac{n}{2}\log(2\pi) \quad \ldots(4.18)$$

$$REML: l_R(\underline{W}, \underline{B}) = -\frac{1}{2}\log|\underline{\Sigma}| - \frac{1}{2}\log|\underline{X}' \underline{\Sigma}^{-1} \underline{X}| - \frac{1}{2}\underline{r}' \underline{\Sigma}^{-1} \underline{r} - \frac{n-p}{2}\log(2\pi) \quad \ldots(4.19)$$

where $\underline{r} = \underline{y} - \underline{X}(\underline{X}' \underline{\Sigma}^{-1} \underline{X})^{-} \underline{X} \underline{\Sigma}^{-1} \underline{y}$ and p is the rank of $\underline{X}$. PROC MIXED procedure of SAS minimizes $-2^* l(\underline{W}, \underline{B})$ or $-2^* l_R(\underline{W}, \underline{B})$ using a ridge-stabilizing Newton-Raphson (NR) algorithm.

4.7.1.2. Estimation of fixed effects parameter vector, $\underline{\beta}$; and random effects parameter vector $\underline{\gamma}$:

Estimates of $\underline{W}$ and $\underline{B}$ obtained by ML or REML are denoted by $\hat{\underline{W}}$ and $\hat{\underline{B}}$ respectively. Estimates of $\underline{\beta}$ and $\underline{\gamma}$ are obtained by solving the mixed model equations (Henderson, 1984):

$$\begin{pmatrix} \underline{X}' \hat{\underline{W}}^{-1} \underline{X} & \underline{X}' \hat{\underline{W}}^{-1} \underline{Z} \\ \underline{Z}' \hat{\underline{W}}^{-1} \underline{X} & \underline{Z}' \hat{\underline{W}}^{-1} \underline{Z} + \underline{B}^{-1} \end{pmatrix} \begin{pmatrix} \hat{\underline{\beta}} \\ \hat{\underline{\gamma}} \end{pmatrix} = \begin{pmatrix} \underline{X}' \hat{\underline{W}}^{-1} \underline{y} \\ \underline{Z}' \hat{\underline{W}}^{-1} \underline{y} \end{pmatrix} \quad \ldots(4.20)$$

The solution of the above equations can be written as:

$$\hat{\underline{\beta}} = (\underline{X}' \underline{\Sigma}^{-1} \underline{X})^{-} \underline{X}' \underline{\Sigma}^{-1} \underline{y} \quad \ldots(4.21)$$

$$\hat{\underline{\gamma}}_i = \underline{B} \underline{Z}' \underline{\Sigma}_i^{-1} (\underline{y} - \underline{X} \hat{\underline{\beta}}) \quad \ldots(4.22)$$

where the superscript (-) in *(4.21)* denotes a generalized inverse. When $\underline{\underline{W}}$ and $\underline{\underline{B}}$ are known, equation *(4.21)* is the best linear unbiased estimator (BLUE) of $\underline{\beta}$ (For definition of BLUE, refer to Appendix A) (Wolfinger et al., 1991) and equation *(4.22)* is the best linear unbiased predictor (BLUP) of $\underline{\gamma}$ (for definition of BLUP, refer to Appendix A) (Wolfinger et al., 1991). The mixed model equation in *(4.20)* assumes that $\hat{\underline{\underline{B}}}$ is nonsingular. If $\hat{\underline{\underline{B}}}$ is singular, then the mixed model equations are modified (Henderson, 1984) as follows:

$$\begin{pmatrix} \underline{X}' \hat{\underline{\underline{W}}}^{-1} \underline{X} & \underline{X}' \hat{\underline{\underline{W}}}^{-1} \underline{Z} \hat{\underline{L}} \\ \hat{\underline{L}}' \underline{Z}' \hat{\underline{\underline{W}}}^{-1} \underline{X} & \hat{\underline{L}}' \underline{Z}' \hat{\underline{\underline{W}}}^{-1} \underline{Z} + \underline{I} \end{pmatrix} \begin{pmatrix} \hat{\underline{\beta}} \\ \hat{\underline{\gamma}} \end{pmatrix} = \begin{pmatrix} \underline{X}' \hat{\underline{\underline{W}}}^{-1} \underline{y} \\ \hat{\underline{L}} \underline{Z}' \hat{\underline{\underline{W}}}^{-1} \underline{y} \end{pmatrix} \qquad \text{... (4.23)}$$

where $\hat{\underline{L}}$ is the lower-triangular Cholesky root of $\hat{\underline{\underline{B}}}$, satisfying $\hat{\underline{\underline{B}}} = \hat{\underline{L}}\,\hat{\underline{L}}'$. The covariance matrix of $\hat{\underline{\beta}}$ and $\hat{\underline{\gamma}}$ are:

$$\underline{\underline{C}} = \begin{bmatrix} \underline{X}' \underline{\underline{W}}^{-1} \underline{X} & \underline{X}' \underline{\underline{W}}^{-1} \underline{Z} \\ \underline{Z}' \underline{\underline{W}}^{-1} \underline{X} & \underline{Z}' \underline{\underline{W}}^{-1} \underline{Z} + \underline{\underline{B}} \end{bmatrix}^{-} \qquad \text{... (4.24)}$$

where $^{-}$ denotes a generalized inverse. However, $\underline{\underline{W}}$ and $\underline{\underline{B}}$ are usually unknown and are estimated by using one of the previously mentioned or other suitable methods. The estimates $\hat{\underline{\underline{W}}}$ and $\hat{\underline{\underline{B}}}$ are therefore substituted in (4.24) to obtain an approximate variance-covariance matrix of $\hat{\underline{\beta}}$, $\hat{\underline{\gamma}}$.

4.7.2. Random effects model for unequally spaced data

A general random effects model (equation 4.16) is described in Section 4.7, Chapter 4. In this model it is assumed that $\underline{\underline{\Sigma}}$ ($= \underline{\underline{Z}}\underline{\underline{B}}\underline{\underline{Z}}' + \underline{\underline{W}}$) and $\underline{\underline{W}}$ is parameterized in terms of a vector of unknown parameters $\underline{\theta}$. For unequally spaced data, one possible

parameterization is a continous-time autoregressive structure with observational error. When the observations are uneually spaced, the AR(1) error structure must be based on a continous-time AR model. A REM was fitted assuming continuous-time/unequally-spaced AR covariance structure, i.e the correlation between any two responses on the same subject equals a base-line correlation value (ρ) raised to a power equal to the absolute difference between the two time points of the responses (defined in Section 4.3.1). An observation error (i.e a matrix $\sigma^2 I$) was added to the continuous-time/unequally-spaced covariance structure and REM was fitted again.

4.8. Goodness of fit statistics for longitudinal models

For Gaussian longitudinal data, the likelihood ratio test and Akaike's information criterion are the two most commonly used goodness-of-fit statistics. For the Gaussian-based linear mixed-effects model, these two methods can be used for assessing the goodness-of-fit for response function and variance-covariance structure. There are two main drawbacks of using likelihood methods: (1) they require complete specification of likelihood function; and (2) they require repeated fittings of the data to a family of nested models. Likelihood based goodness-of-fit measures can not be used to assess the adequacy of models which are fitted by using generalized estimating equations. Vonesh et al. (1996) provided a measure of concordance between fitted and observed responses which is similar to the coefficient of determination, R^2, used in univariate linear regression settings. In addition, they provided a measure of concordance between assumed and true covariance structure, and the pseudo likelihood

ratio test to test the null hypothesis that assumed and true covariance structures are equal.

The likelihood ratio test is described in section 4.8.1 and Akaike's information criterion is described in section 4.8.2. The concordance correlation coefficient for response function is given in section 4.8.3; the concordance correlation coefficient to measure the closeness between assumed and covariance structure and pseudo likelihood ratio test is given in section 4.8.4.

4.8.1. Likelihood ratio test:

The likelihood ratio test is used to determine the goodness-of-fit of a model. When two models are fitted to the same data by the maximum likelihood method, and one model is a constrained version of the other model, then the likelihood ratio test can be used to test the following null hypothesis:

H_o: The model with more parameters is not a significantly better model than the model with fewer parameters.

The model with more parameters usually has all the parameters of the constraint model and a few extra parameters. The following steps are involved in likelihood ratio test:

Step 1.: Fit a model with a certain number of parameters, which includes both linear and non-linear parameters.

Step 2: Fit another model which has the same parameters as in Step 1 plus k extra parameters.

Let L_1 be the value of -2 log-likelihood from the first model and L_2 be the -2 log-likelihood from the second model with k extra parameters, the likelihood ratio test

states that under the null hypothesis the two models are the same, i.e extra parameters are zero. Change in -2 log-likelihood for these two models is asymptotically distributed as χ^2 with k degrees of freedom, which is the difference in the number of parameters in the two models; i.e.

$$L_1 - L_2 \sim \chi^2_k \qquad \ldots(4.25)$$

A large value of chi-square is in the direction of rejecting the null hypothesis that extra k parameters are zero.

The likelihood ratio test cannot be used in all situations. If we have two models with one a constrained or reduced version of the other, then the two such models are called "nested". When we have to fit a number of models, the likelihood ratio test can be calculated for any two nested models, by calculating the difference of their -2 log-likelihoods, but there will be several pairs which are not nested, so the likelihood ratio test can not be used. In later situations, or in general for overall model selection, Akaike's Information Criterion is very useful.

4.8.2. Akaike's information criterion (AIC):

Akaike's Information Criterion (AIC) (Akaike, 1973, 1974) is another method which helps us with the model selection. AIC is based on decision theory. Akaike developed the AIC criterion for the identification of an optimal and parsimonious model in data analysis from a class of competing models which takes model complexity into account. It penalizes -2 log-likelihood for the number of fitted parameters to the data to avoid overfitting. AIC is an extension of the classical maximum likelihood principle. It

combines the maximum likelihood and the corresponding log-likelihood ratio statistics, i.e.

AIC = -2 log(maximum likelihood) + 2 (number of estimated parameters)...(4.26).

In particular, in longitudinal analysis, we select the most appropriate model by choosing one with the minimum AIC value. Therefore, the model with the lowest AIC is chosen as the 'best' model.

4.8.3. Concordance correlation coefficient for response function

For longitudinal studies, let y_{ij} be the j^{th} observed response for i^{th} subject and $\hat{y}_{ij}$ be the predicted response, Vonesh et al (1996), defined the model concordance correlation coefficient as:

$$r_c = 1 - \frac{\Sigma(y_i - \hat{y}_i)'(y_i - \hat{y}_i)}{\Sigma(y_i - \bar{y}1_i)'(y_i - \bar{y}1_i) + \Sigma(\hat{y}_i - \hat{y}1_i)'(\hat{y}_i - \hat{y}1_i) + N(\bar{y} - \hat{y})} \qquad ...(4.27)$$

where $\bar{y}$ is the grand mean of y_{ij} and $\hat{y}$ is the grand mean of $\hat{y}_{ij}$, and 1_i is the $n_i \times 1$ vector of ones (n_i is the number of observations for the i^{th} subject) and $N = \Sigma n_i$ is the total number of observations. The advantages of r_c over the traditional coefficient of determination, R^2, in multiple regression setting are (Vonesh et al., 1996):

- r_c is directly interpretable as a concordance correlation between observed and predicted values and directly measures the level of agreement between observed y_i and predicted $\hat{y}_i$. Its value shows how well a scatter plot of y_{ij} versus $\hat{y}_{ij}$ about the line of identity.
- The line of identity serves as a reference so it does not require specification of a null model.

- The possible values of r_c lie between -1 and +1. The value of r_c equal to one implies the perfect fit and the value of r_c less than or equal to zero implies the lack of fit.

The Pearson correlation coefficient for simple linear regression will be same for intercept or no-intercept model, while the concordance coefficient distinguishes between the intercept and no-intercept model (Vonesh et al., 1996). The value of r_c behaves like R^2, so it will increase with more complicated or over parameterized models. Similar to R^2, we can adjust r_c for a number of parameters. The adjusted value is given by $r_{c,a} = 1 - k(1 - r_c)$. Here $k = N/(N - p)$ is a correction factor adjusting for number of parameters (p) in β.

4.8.4 Goodness-of-fit statistic for variance-covariance structure

The utility of any statistical model is determined by its ability to approximate the unknown response and as well as by its ability to account for "background noise". Vonesh et al (1996) also defined a goodness-of-fit statistic for assessing the adequacy of the variance-covariance structure given that the underlying response function has been correctly specified. The closeness between the assumed covariance structure (Σ_i) and the unknown true covariance structure $[var(y_i)]$ is measured by the variance-covariance concordance correlation co-efficient, $r(\hat{\omega})$. Our aim is to evaluate whether the unknown true covariance structure is well approximated by the assumed variance covariance structure or not. For a balanced and complete set of data, the within subject covariance structure is the same for all subjects, (i.e $\Sigma_i = \Sigma$). In such cases, one can compare the unspecified covariance structure to the assumed/specified covariance structure using the likelihood ratio test. This test assumes that y_i's are normally

distributed. One can use the likelihood ratio test for slightly unbalanced data sets. In this approach, one has to fit the data twice, one with unspecified covariance structure and one with the assumed covariance structure. This approach fails when data are highly unbalanced and the likelihood method is not used for estimation. Vonesh et al. (1996) resolved this problem by proposing a concordance correlation coefficient for measuring the closeness between assumed and true covariance structures and a pseudo-likelihood ratio test for testing the equality between true and assumed covariance structure. These new goodness-of-fit statistics are explained below.

Under suitable regularity conditions, Liang and Zeger (1986) showed $\hat{\underline{\beta}}$ is a consistent estimate of $\underline{\beta}$ as n $\rightarrow \infty$, and that $\sqrt{m}(\hat{\underline{\beta}}-\underline{\beta})$ is asymptotically distributed with mean $\underline{0}$ and variance-covariance structure:

$$\text{(i)} \qquad \underline{\Omega}_1 = \lim_{m\to\infty}\left\{ m\underline{\Omega}\left(\sum_{i=1}^{m}\left(\frac{\partial \underline{\mu}_i}{\partial \underline{\beta}}\right)' \underline{\Sigma}_i^{-1}\,\mathrm{var}(\underline{y}_i)\underline{\Sigma}_i^{-1}\left(\frac{\partial \underline{\mu}_i}{\partial \underline{\beta}}\right)\right)\underline{\Omega}\right\} \qquad \ldots (4.28)$$

$$\text{where} \qquad \underline{\Omega} = \left[\sum_{i=1}^{m}\left(\frac{\partial \underline{\mu}_i}{\partial \underline{\beta}}\right)' \underline{\Sigma}_i^{-1}\left(\frac{\partial \underline{\mu}_i}{\partial \underline{\beta}}\right)\right]^{-1} \qquad \ldots (4.29)$$

and

$$\text{(ii)} \qquad \underline{\Omega}_0 = \lim_{m\to\infty}\{m\underline{\Omega}\} \qquad \ldots (4.30)$$

by assuming $var(y_i) = \underline{\Sigma}_i$.

To assess the goodness of fit of assumed variance-covariance structure, it is important to assume that the response function is correctly specified. If in addition, the

variance-covariance matrix of $\underline{\Sigma}_i$ is correctly specified, then inference on $\underline{\beta}$ can be carried out on the basis that $\sqrt{m}(\hat{\underline{\beta}}-\underline{\beta})$ will be asymptotically normally distributed with the variance-covariance matrix consistently estimated by:

$$\hat{\text{var}}(\hat{\underline{\beta}}) = \hat{\underline{\Omega}} = \left[\sum_{i=1}^{m}\left(\frac{\partial \hat{\underline{\mu}}_i}{\partial \underline{\beta}}\right)'(\underline{\Sigma}_i))^{-1}\left(\frac{\partial \hat{\underline{\mu}}_i}{\partial \underline{\beta}}\right)\right]^{-1} \quad \text{.... (4.31)}$$

Alternatively, for generalized estimation equations, Liang and Zeger (1986) recommended the use of the following robust estimator:

$$\hat{\underline{\Omega}}_R = \hat{\text{var}}(\hat{\underline{\beta}}) = \hat{\underline{\Omega}}^{-1}\,\hat{\underline{\Omega}}_1\,\hat{\underline{\Omega}}^{-1} \qquad \text{... (4.32)}$$

where $\hat{\underline{\Omega}}$ is as in equation *(4.31)* and

$$\hat{\underline{\Omega}}_1^{-1} = \sum_{i=1}^{m}\left(\frac{\partial \hat{\underline{\mu}}_i}{\partial \underline{\beta}}\right)'(\underline{\Sigma}_i)^{-1}(\underline{y}_i - \hat{\underline{\mu}}_i)(\underline{y}_i - \hat{\underline{\mu}}_i)'(\underline{\Sigma}_i)^{-1}\left(\frac{\partial \hat{\underline{\mu}}_i}{\partial \underline{\beta}}\right) \text{.... (4.33)}$$

in the above equation $(\underline{y}_i - \hat{\underline{\mu}}_i)(\underline{y}_i - \hat{\underline{\mu}}_i)'$ is the estimated covariance matrix of residuals, where $\hat{\underline{\mu}}_i$ is estimated using the inverse of the link function $g(\hat{\mu}_{ij}) = \underline{X}_{ij}\,\hat{\underline{\beta}}$.

The advantage of using the robust estimate given in *(4.32)* is that it provides $\sqrt{m}$-consistent estimate of the covariance matrix of $\hat{\underline{\beta}}$ even under the misspecification of covariance structure provided the marginal response function is specified correctly. From equations *(4.31)* and *(4.32)*, $m\hat{\Omega}_R$ is a consistent estimate of $\underline{\Omega}_1$ and $m\hat{\Omega}$ is a consistent estimate of $\underline{\Omega}_0$. If the covariance structure is correctly specified then use of $\hat{\Omega}_R$ may be less efficient than $\hat{\Omega}$ (Vonesh et al., 1996). The use of $\hat{\Omega}_R$ in combination with a misspecified covariance matrix can result in a significant loss in efficiency as

compared to the use of either $\hat{\Omega}$ or $\hat{\Omega}_R$ under the correct specification of Σ_i (Vonesh et al , 1996). For this reason, we need goodness-of-fit statistics for selecting not only a correct response function but also for selecting an appropriate covariance structure. Vonesh et al (1996) resolved this problem by proposing a goodness-of-fit statistic based on comparing the robust "sandwich" estimator of $\hat{\beta}$ ($\hat{\underline{\Omega}}_R$) and the estimated asymptotic covariance matrix of $\hat{\beta}$ computed under the assumed covariance structure ($\hat{\underline{\Omega}}$) using quasi-likelihood methods.

From equations *(4.28)* and *(4.30)*, we see whenever $var(y_i) = \Sigma_i$, then $\underline{\Omega}_1 = \underline{\Omega}_0$. So, in order to assess the goodness-of-fit of Σ_i to $var(y_i)$, Vonesh et al (1996) recommended the comparison of $\hat{\underline{\Omega}}_R$ to $\hat{\underline{\Omega}}$. Vonesh et al. (1996) gave a descriptive statistic to measure the closeness of $\hat{\underline{\Omega}}_R$ and $\hat{\underline{\Omega}}$, and a pseudo-likelihood ratio test for testing the equality of $\hat{\underline{\Omega}}_R$ and $\hat{\underline{\Omega}}$.

The variance-covariance concordance correlation coefficient, $r(\hat{\omega})$, measures the closeness between $\hat{\underline{\Omega}}_R$, the robust "sandwich" estimator of the covariance of $\hat{\beta}$, with $\hat{\underline{\Omega}}$, the estimated asymptotic covariance matrix of $\hat{\beta}$ computed under the assumed covariance structure.

Let $\hat{\omega} = vech\{(\hat{\underline{\Omega}})^{-1/2}\,\hat{\underline{\Omega}}_R\,(\hat{\underline{\Omega}})^{-1/2}\}$ be the vector formed from the lower triangular part of the unitless matrix $\{(\hat{\underline{\Omega}})^{-1/2}\,\hat{\underline{\Omega}}_R(\hat{\underline{\Omega}})^{-1/2}\}$. Let i be the vector formed from the lower triangular part of the $p \times p$ identity matrix I_p. Vonesh et al. (1996) proposed the variance-covariance concordance correlation to be:

$$r(\hat{\omega}) = 1 - \frac{\|(\hat{\omega} - \hat{\omega})\|^2}{\|\hat{\omega}\|^2 + \|\hat{\omega}\|^2} = 1 - \frac{\|(\hat{\omega} - \hat{\omega})\|^2}{\|\hat{\omega}\|^2 + p} \qquad ...(4.34)$$

where $\|x\|$ is the Euclidian norm of a vector $\underline{x}$. The norm of vector $\underline{x}$ is defined by: $\|x\| = (\underline{x}.\underline{x})^{1/2} = (\Sigma x_i^2)^{1/2}$.

The variance-covariance concordance correlation coefficient, $r(\hat{\omega})$ is useful in detecting gross differences between the assumed covariance structure, $\underline{\Sigma}_i$, and the actual structure, var($\underline{y}_i$). Vonesh et al (1996) found that it is less sensitive to moderate differences.

4.8.5 Pseudo-likelihood ratio test

To assess the goodness-of-fit statistic for a specified covariance structure, we have to assume that the underlying response function is correctly specified. Vonesh et al (1996) proposed an alternative test called a pseudo-likelihood test for testing the equality of two covariance structures. The null hypothesis $H_0 : var(\underline{y}_i) = \underline{\Sigma}_i$ will hold true provided the hypothesis $H_0 : \underline{\Omega}_1 = \underline{\Omega}_0$ holds true. The pseudo-likelihood test is given by :

$$\hat{\lambda} = m\{ln|\underline{\hat{\Omega}}| - ln|\underline{\hat{\Omega}}_R| + trace(\underline{\hat{\Omega}}_R \underline{\hat{\Omega}}^{-1}) - p\} \qquad ...(4.35)$$

Vonesh et al (1996) showed that under $H_0 : \underline{\Omega}_1 = \underline{\Omega}_0$, the statistic is approximately distributed as chi-square with $p(p+1)/2$ degrees of freedom provided $m\underline{\hat{\Omega}}_R$ has an approximate Wishart distribution, $W_s(m, \underline{\Omega}_i/m)$. The assumption that $m\underline{\hat{\Omega}}_R$ has an approximate Wishart distribution seems reasonable provided the $\underline{y}_i$ are approximately normally distributed. The pseudo-likelihood ratio test performs well under conditions of

normality but it is sensitive to departures from normality. Vonesh et al (1996) applied this test to real data sets to illustrate the utility of this test for the continuous and discrete data. Vonesh et al (1996) also presented the results from a limited simulation study designed to evaluate the robustness of the pseudo-likelihood ratio test to non-normally distributed data. Their results showed that under normality, the pseudo-likelihood ratio test compares favourably with the standard normal-theory likelihood ratio test, but it is sensitive to departures from normality. Further applications of Vonesh's goodness-of-fit statistics to small samples and non-gaussian data are required to examine the behaviour of these statistics.

5. RESULTS – STATISTICAL MODELS FOR LONGITUDINAL PULMONARY FUNCTION DATA ON GRAIN ELEVATOR WORKERS

5.1 Introduction

Our objective was to explore the role played by different longitudinal models in predicting longitudinal decline in the lung function measures (FEV_1 and FVC) of grain elevator workers. This was conducted by fitting different models; namely marginal, transitional, and random effects in the context of predicting longitudinal changes in lung function. We assessed the adequacy of the models and the assumed variance-covariance structure by using the model concordance coefficient; variance-covariance concordance coefficient and pseudo-likelihood ratio test respectively.

5.2. Statistical methods

The data used in this chapter was a part of the Labour Canada GDMSP which was described in Section 3.2, Chapter 3. Marginal and transitional models were fitted by using generalized estimating equations. The random effects model was fitted using the maximum likelihood method. The four steps involved in the longitudinal data analysis were: (i) the choice of the model; (ii) the choice of the variance-covariance structure; (iii) assessing the goodness-of-fit of the model; and (iv) assessing the goodness-of-fit of the variance covariance structure.

Marginal, transitional, and random effects models were fitted assuming the following within-subject correlation structures:

1. Uncorrelated covariance structure, i.e. the longitudinal observations for a subject are not correlated;
2. Compound symmetric; i.e off diagonal elements of the covariance matrix are all the same;
3. Unspecified correlation structure; i.e all the elements of covariance are to be estimated.
4. Autoregressive covariance structure assuming the observations within subject are equally spaced; i.e the correlation between any two responses on the same subject equals a baseline correlation value (ρ) raised to a power equal to the difference between the two time points. The time points were assumed to be equally spaced ($t_{i1}=1$; $t_{i2}=2$, $t_{i3}=3$ and so on).

Random effects models were also fitted assuming unequally-spaced AR(1) covariance structure, i.e the correlation between any two responses on the same subject equals a base-line correlation value (ρ) raised to a power equal to the absolute difference between the two time points of the responses. An observational error (i.e a matrix $\sigma^2 I$) was added to the unequally-spaced AR(1) covariance structure and REM was fitted again.

5.2.1 Marginal model

Marginal models were described in Section 4.5, Chapter 4. The covariates considered in the marginal model were: height at baseline; weight; age; years in the grain industry; two dummy variables for smoking (exsmokers and smokers); four dummy variables for cycle (Cycle II, Cycle III, Cycle IV, and Cycle V); an interaction

term between years in the grain industry and exsmokers; and an interaction term between years in the grain industry and current smokers. The marginal model is given by:

$(FEV_1)_{ij} = \beta_0 + \beta_1*(Base\ height)_i + \beta_2*(Weight)_{ij} + \beta_3*(Age)_{ij} + \beta_4*(Yrs.\ in\ grain\ industry)_{ij} + \beta_5*(Exsmoker)_{ij} + \beta_6*(Smokers)_{ij} + \beta_7*(Cycle\ II)_{ij} + \beta_8*(Cycle\ III)_{ij} + \beta_9*(Cycle\ IV)_{ij} + \beta_{10}*(Cycle\ V)_{ij} + \beta_{11}*(Yrs.\ in\ grain\ industry*exsmokers)_{ij} + \beta_{12}*(Yrs.\ in\ grain\ industry*smokers)_{ij} + \epsilon_{ij}\ ;\ i = 1, 2, ..., n\ ;\ j = 1,2, ..., N_i$ ... (5.1)

where $(FEV_1)_{ij}$, $(Weight)_{ij}$, $(Age)_{ij}$, $(Yrs.\ in\ grain\ industry)_{ij}$ denote the measurements for i^{th} grain worker at time j.

$(Exsmokers)_{ij}$ and $(Smoker)_{ij}$ are binary variables defined as:

$(Exsmoker)_{ij}$ = 1 if i^{th} grain worker is exsmoker at time j.

= 0 otherwise

$(Smoker)_{ij}$ = 1 if i^{th} grain worker is smoker at time j.

= 0 otherwise

In matrix notation, marginal model can be written as:

$$\underline{Y} = \underline{X}\underline{\beta} + \underline{\epsilon} \qquad \text{.... (5.2)}$$

where $\underline{Y}$ is a $(\sum_{i=1}^{n} N_i)$x1 response vector; $\underline{X}$ is a known $(\sum_{i=1}^{n} N_i)$x13 design matrix; $\underline{\beta}$ is a 13x1 vector of unknown parameters; and $\underline{\epsilon}$ is a $(\sum_{i=1}^{n} N_i)$x1 an error vector. Marginal models were also fitted without age in the model.

5.2.2 Transitional model

Transitional models were described in Section 4.6, Chapter 4. In addition to the covariates used in the marginal model, a covariate was included to indicate the previous lung function in the transitional model.

$$(FEV_1)_{ij} = \beta_0 + \beta_1*(FEV_1)_{ij-1} + \beta_2*(Base\ height)_i + \beta_3*(Weight)_{ij} + \beta_4*(Age)_{ij} + \beta_5*(Yrs.\ in\ grain\ industry)_{ij} + \beta_6*(Exsmoker)_{ij} + \beta_7*(Smokers)_{ij} + \beta_8*(Cycle\ II)_{ij} + \beta_9*(Cycle\ III)_{ij} + \beta_{10}*(Cycle\ IV)_{ij} + \beta_{11}*(Cycle\ V)_{ij} + \beta_{12}*(Yrs.\ in\ grain\ industry*exsmokers)_{ij} + \beta_{13}*(Yrs.\ in\ grain\ industry*smokers)_{ij} + \epsilon_{ij} \quad ...(5.2)$$

where $\underline{Y}$ is a $(\sum_{i=1}^{n} N_i)$x1 response vector; $\underline{X}$ is a known $(\sum_{i=1}^{n} N_i)$x14 design matrix; $\underline{\beta}$ is a 14x1 vector of unknown parameters; and $\underline{\epsilon}$ is a $(\sum_{i=1}^{n} N_i)$x1 an error vector.

Transitional models were also fitted without age in the model.

5.2.3 Random effects model

Random effects models were described in Section 4.7, Chapter 4. In addition to the covariates used in the marginal model, a variable was included to indicate the random effect of the variable age.

$$(FEV_1)_{ij} = \beta_0 + \beta_1*(Base\ height)_i + \beta_2*(Weight)_{ij} + \beta_3*(Age)_{ij} + \beta_4*(Yrs.\ in\ grain\ industry)_{ij} + \beta_5*(Exsmoker)_{ij} + \beta_6*(Smokers)_{ij} + \beta_7*(Cycle\ II)_{ij} + \beta_8*(Cycle\ III)_{ij} + \beta_9*(Cycle\ IV)_{ij} + \beta_{10}*(Cycle\ V)_{ij} + \beta_{11}*(Yrs.\ in\ grain\ industry*exsmokers)_{ij} + \beta_{12}*(Yrs.\ in\ grain\ industry*smokers)_{ij} + \gamma_i(age)_{ij} + \epsilon_{ij} \quad .... (5.3)$$

where γ_i denotes the random effect of age. In matrix notation, a random effects model can be written as :

$$\underline{Y} = \underline{X}\underline{\beta} + \underline{Z}\gamma_i + \underline{\epsilon} \qquad (5.4)$$

where $\underline{Y}$ is a response vector, $\underline{X}$ is a design matrix for fixed effects and $\underline{Z}$ is a vector, which contains values of age. Vector $\underline{\beta}$ is a *13x1* vector of parameters β_0 to β_{12} defined in the marginal model. A random effect γ_i is distributed as N(0,b), $\underline{\epsilon}$ is distributed as N(0, $\underline{W}$), γ_i and $\underline{\epsilon}$ are independently distributed. Random effects models were also fitted without the fixed effect parameter for age.

Random effects models were fitted assuming the four different within-subject covariance structures mentioned in Section 5.2, which assumed that the data were equally spaced. A random effects model was also fitted for unequally spaced data (Section 4.7.2, Chapter 4). In the Grain Dust Medical Surveillance Program (Section 3.1, Chapter 3), the underlying basic sampling interval was three years, i.e observations on grain workers were collected every three years. Even though the intended time interval was three years, the time between any two observations was not exactly three years, i.e observations were not equally spaced. Therefore, the data were treated as unequally-spaced. Information on the date of examination was used to calculate the exact sampling interval for each grain elevator worker. When the observations are unequally spaced, the AR(1) (autoregressive) error structure must be based on the continuous-time/unequally-spaced AR model. For random effects models, model and variance-covariance concordance coefficients and pseudo-likelihood ratio tests were computed and compared with the traditional likelihood ratio test and Akaike's

information criterion, which are generally used for assessing the goodness of fit of models based on maximum/restricted maximum likelihood theory.

Initially, two random effects models were considered; random intercept and random slope. Due to the large sample size, the maximum virtual memory required to run the SAS program was greater than that was available on the VMS 6.2 operating system on the mainframe computer. Therefore, a model with a single random effect was considered with the assumption that each individual has the same intercept as the population intercept.

Marginal and transitional models were fitted using the generalized estimating equations approach using a SAS MACRO developed by Karim (1986). PROC MIXED(1994) procedure in SAS was used to fit the random effects model.

5.2.4 Goodness of fit statistics

In this chapter, we used the model and variance-covariance concordance correlation coefficients and pseudo-likelihood ratio test to assess the goodness of fit of model and variance covariance structure. These goodness of fit statistics are explained in detail in Section 4.8 of Chapter 4.

5.3. Results

5.3.1. Subjects

The number of grain workers in the province of Saskatchewan, who participated in the Medical Surveillance Program are given in Table 5.1. A further breakdown of these grain workers by their participation in different cycles is given in Table 5.2. There

were 203 grain workers who participated in all five cycles and contributed 1015 observations; 259 grain workers who participated in any four cycles and contributed 1036 observations; 497 grain workers who participated in any three cycles and contributed 1491 observations; 739 workers participated in any two cycles and contributed 1478 observations; and grain workers who participated only in any one cycle contributed 2394 observations (Table 5.3).

Table 5.1: Number of grain workers in each Cycle

	No. of grain workers
Cycle I	2551
Cycle II	1930
Cycle III	883
Cycle IV	1005
Cycle V	1045

5.3.2. Descriptive statistics

Table 5.4 shows the mean demographic, smoking and lung function test values at each of the five cycles. The mean age of grain workers at Cycle I was 33.14 yrs and 36.47 yrs. at Cycle V, the difference of 3.34 years in mean age over 15 years indicates that older grain workers either retired or quit the grain industry and younger workers joined the grain industry. Mean years in grain industry was 14.03 at Cycle V and 8.74 at Cycle I, which again indicates that young grain workers joined the grain

industry and older workers had quit the industry. Tables 5.5 and 5.6 show the mean ± standard deviation for FEV_1 and FVC stratified by the presence and absence of symptoms (cough, sputum, wheeze, or dyspnea) respectively. Grain workers with no symptoms had significantly higher FEV_1 compared to those who reported any symptom (Table 5.5). We observed similar results for FVC (Table 5.6). Table 5.7 shows the comparison of lung function measurements by smoking status. Smokers and ex-smokers had significantly lower FEV_1 and FVC compared to non-smokers.

5.3.3 Univariate regression analysis

Univariate regression analysis for each cycle separately showed that age; height and years in the grain industry were significantly associated with FEV_1 when each of these factors was introduced one at a time in the regression model (Table 5.8). When these models were fitted for FVC, in addition to the above mentioned factors, weight was also significantly associated with FVC (Table 5.9). Based on these results, the potential predictors for the multivariate longitudinal model for predicting the lung function variable (FEV_1 or FVC) were age, height, weight, years in the industry, smoking status, survey effect. We used base height in our model, while other covariates were used as time-varying.

Table 5.2 : Pattern of participation of grain elevator workers in Saskatchewan over 15 years (1978 - 1993)

Cycle Number					Number of grain workers
I	II	III	IV	V	
Observations with base-line Cycle I:					
X	X	X	X	X	203
X	X	X	X	-	98
X	X	X	-	X	74
X	X	-	X	X	36
X	-	X	X	X	12
X	X	X	-	-	181
X	X	-	X	-	122
X	-	X	X	-	11
X	X	-	-	X	78
X	-	X	-	X	6
X	-	-	X	X	1
X	X	-	-	-	303
X	-	-	X	-	10
X	-	X	-	-	14
X	-	-	-	X	3
X	-	-	-	-	1399
Observations with base-line Cycle II:					
-	X	X	X	X	39
-	X	X	X	-	28
-	X	X	-	X	19
-	X	-	X	X	36
-	X	-	X	-	74
-	X	X	-	-	67
-	X	-	-	X	187
-	X	-	-	-	385
Observations with base-line Cycle III:					
-	-	X	X	X	15
-	-	X	X	-	25
-	-	X	-	X	6
-	-	X	-	-	85
Observations with base-line Cycle V:					
-	-	-	X	X	50
-	-	-	X	-	245
Observations at Cycle V:					
-	-	-	-	X	280

Table 5.3 : Number of observation available on grain workers

	No. of observations				
	5	4	3	2	1
No. of grain workers	203	259	497	739	2394
Total no. of observations	1015	1036	1491	1478	2394

Table 5.4 : Descriptive statistics of demographic variables by Cycle.

	Cycle I Mean ± S.D	Cycle II Mean ± S.D	Cycle III Mean ± S.D	Cycle IV Mean ± S.D	Cycle V Mean ± S.D
Age, yrs	33.14 ± 13.57	34.48 ± 12.89	35.50 ± 2.49	35.64 ± 0.73	36.47 ± 9.99
Height, cm	174.39 ± 6.36	175.97 ± 6.21	174.86 ± 5.90	176.02 ± 6.28	176.24 ± 6.79
Weight, Kg	81.88 ± 13.61	83.53 ± 13.77	84.89 ± 13.34	83.95 ± 13.95	85.18 ± 13.67
Yrs. In grain industry	8.74 ± 10.11	11.39 ± 9.79	12.62 ± 9.81	12.91 ± 8.81	14.03 ± 8.33
FEV_1, litres	4.04 ± 0.79	4.19 ± 0.72	4.01 ± 0.72	4.20 ± 0.71	4.25 ± 0.73
FVC, litres	5.22 ± 0.94	5.37 ± 0.80	5.11 ± 0.81	5.32 ± 0.83	5.34 ± 0.84

Table 5.5 : Mean and standard deviation for FEV_1 (liters) by symptoms

		Cycle I Mean ± S.D	Cycle II Mean ± S.D	Cycle III Mean ± S.D	Cycle IV Mean ± S.D	Cycle V Mean ± S.D
Cough	Yes	3.85 ± 0.87***	4.10 ± 0.77*	3.73 ± 0.75****	3.93 ± 0.75****	4.10 ± 0.84*
	No	4.13 ± 0.76	4.21 ± 0.70	4.04 ± 0.71	4.24 ± 0.69	4.27 ± 0.73
Wheeze	Yes	3.79 ± 0.86***	4.01 ±0.74****	3.87 ± 0.75**	3.77 ± 0.80****	3.82 ± 0.85***
	No	4.13 ± 0.77	4.24 ± 0.71	4.05 ± 0.71	4.22 ± 0.69	4.28 ± 0.72
Sputum	Yes	3.80 ± 0.87***	4.04 ± 0.83***	3.73 ± 0.77****	3.96 ± 0.80**	4.11 ± 0.78
	No	4.13 ± 0.76	4.23 ± 0.69	4.04 ± 0.71	4.22 ± 0.69	4.27 ± 0.73
Dyspnea	Yes	3.49 ± 0.83***	3.67 ±0.75****	3.45 ± 0.78****	3.76 ± 0.64****	3.85 ±0.63**
	No	4.17 ± 0.74	4.26 ± 0.69	4.08 ± 0.68	4.23 ± 0.70	4.26 ± 0.74

*: <0.05
**: < 0.01
***: < 0.001
****: < 0.0001

Table 5.6 : Mean and standard deviation for FVC (liters) by symptoms

		Cycle I Mean ± S.D	Cycle II Mean ± S.D	Cycle III Mean ± S.D	Cycle IV Mean ± S.D	Cycle V Mean ± S.D
Cough	Yes	5.15 ± 1.00***	5.40 ± 0.82	4.93 ± 0.86*	5.17 ± 0.87*	5.33 ± 0.88
	No	5.32 ± 0.90	5.37 ± 0.80	5.13 ± 0.79	5.34 ± 0.81	5.35 ± 0.84
Wheeze	Yes	5.18 ± 0.99*	5.31 ± 0.80	5.01 ± 0.83*	5.06 ± 0.78*	5.01 ± 0.91**
	No	5.30 ± 0.92	5.40 ± 0.81	5.15 ± 0.79	5.33 ± 0.82	5.36 ± 0.83
Sputum	Yes	5.11 ± 0.95****	5.36 ± 0.87	4.96 ± 0.85*	5.20 ± 0.92	5.29 ± 0.87
	No	5.33 ± 0.92	5.38 ± 0.80	5.13 ± 0.79	5.33 ± 0.81	5.35 ± 0.84
Dyspnea	Yes	4.88 ± 1.02***	5.04 ± 0.77***	4.64 ± 0.82****	4.96 ± 0.71***	5.00 ± 0.70*
	No	5.36 ± 0.89	5.42 ± 0.80	5.17 ± 0.78	5.34 ± 0.83	5.35 ± 0.84

*: <0.05
**: < 0.01
***: < 0.001
****: < 0.0001

Table 5.7 : Comparison of lung function measurements by smoking status.

		Cycle I Mean ± S.D	Cycle II Mean ± S.D	Cycle III Mean ± S.D	Cycle IV Mean ± S.D	Cycle V Mean ± S.D
FEV_1						
	Non smokers	4.25 ± .73***	4.38 ± 0.65***	4.25 ± 0.61***	4.40 ± 0.65***	4.42 ± 0.69***
	Ex smokers	3.79 ± .84	4.03 ± 0.75	3.81 ± 0.77	4.07 ± 0.75	4.10 ± 0.76
	Current smokers	4.02 ± .77	4.17 ± 0.71	3.96 ± 0.72	4.13 ± 0.70	4.20 ± 0.73
FVC						
	Non smokers	5.32 ± .84***	5.43 ± 0.80***	5.27 ± 0.76***	5.44 ± 0.79***	5.46 ± 0.84***
	Ex smokers	5.09 ± .99	5.27 ± 0.81	4.96 ± 0.83	5.22 ± 0.86	5.17 ± 0.84
	Current smokers	5.21 ± .91	5.38 ±0.80	5.10 ± 0.80	5.30 ± 0.82	5.36 ± 0.81

*: <0.05
**: < 0.01
***: < 0.001
****: < 0.0001

Table 5.8 : Results of univariate regression showing the factors associated with FEV_1

Model	Cycle I	Cycle II	Cycle III	Cycle IV	Cycle V
	$\hat{\beta}_1$ *(s.e of $\hat{\beta}_1$)*	$\hat{\beta}_1$ *(s.e of $\hat{\beta}_1$)*	$\hat{\beta}_1$ *(s.e of $\hat{\beta}_1$)*	$\hat{\beta}_1$ *(s.e of $\hat{\beta}_1$)*	$\hat{\beta}_1$ *(s.e of $\hat{\beta}_1$)*
$FEV_1=\beta_0+\beta_1 age$	-0.0323*** (0.0010)	-0.0311*** (0.0011)	-0.0363*** (0.0015)	-0.0339*** (0.0018)	-0.0359*** (0.0020)
$FEV_1=\beta_0+\beta_1 height$	0.0341*** (0.0024)	0.0368*** (0.0024)	0.0404*** (0.0037)	0.0446*** (0.0033)	0.0430*** (0.0031)
$FEV_1=\beta_0+\beta_1 weight$	0.00004 (0.0012)	0.0001 (0.0012)	0.0026 (0.0018)	0.0009 (0.0016)	0.0031 (0.0017)
$FEV_1=\beta_0+\beta_1$(yrs. In industry)	-0.0402*** (0.0023)	-0.0371*** (0.0015)	-0.0412*** (0.0020)	-0.0349*** (0.0023)	-0.0324*** (0.0026)

*: <0.05
**: < 0.01
***: < 0.001
****: < 0.0001

Table 5.9 : Results of univariate regression showing the factors associated with dependent variable FVC

Model	Cycle I	Cycle II	Cycle III	Cycle IV	Cycle V
	$\hat{\beta}_1$ *(s.e of $\hat{\beta}_1$)*	$\hat{\beta}_1$ *(s.e of $\hat{\beta}_1$)*	$\hat{\beta}_1$ *(s.e of $\hat{\beta}_1$)*	$\hat{\beta}_1$ *(s.e of $\hat{\beta}_1$)*	$\hat{\beta}_1$ *(s.e of $\hat{\beta}_1$)*
$FVC = \beta_0 + \beta_1 age$	-0.00193*** (0.0013)	-0.0200*** (0.0013)	-0.0306*** (0.0019)	-0.0271*** (0.0023)	-0.0328*** (0.0024)
$FVC = \beta_0 + \beta_1 height$	0.0418*** (0.0028)	0.0543*** (0.0026)	0.0604*** (0.0039)	0.0691*** (0.0035)	0.0629*** (0.0033)
$FVC = \beta_0 + \beta_1 weight$	0.0080*** (0.0014)	0.0060*** (0.0013)	0.0067*** (0.0020)	0.0048* (0.0019)	0.0064*** (0.0019)
$FVC = \beta_0 + \beta_1$(yrs. In industry)	-0.0246*** (0.0029)	-0.0235*** (0.0018)	-0.0336*** (0.0025)	-0.0262*** (0.0029)	-0.0292*** (0.0030)

*: <0.05
**: < 0.01
***: < 0.001
****: <0.0001

5.3.4 Marginal, transitional, and random effects model assuming equally spaced data

Data from Saskatchewan were used to fit marginal, transitional and random effects models. We attempted to fit 48 models; 24 models for the prediction of FEV_1 and another 24 models for the prediction of FVC. There were some numerical problems related to convergence for four transitional models with unspecified covariance structure. We had complete results for 44 models; 8 marginal; 6 transitional; and 8 random effects for the prediction of FEV_1 and the same number of models for the prediction of FVC. In these models we used two dummy variables for smoking status (these compared exsmokers vs non-smokers; and current smokers vs. non-smokers); four dummy variables to study the effect of cycle (the reference category was Cycle I); two interaction terms for years in the grain industry and exsmokers and years in the the grain industry and current smokers. The transitional model for the prediction of FEV_1 (FVC) also had previous FEV_1 (previous FVC) in the model in addition to the above mentioned independent variables. In the random effects model, parameter associated with age was considered as a random effect (see Section 5.2.3).

5.3.4.1 Marginal, transitional and random effects models for FEV_1 assuming equally spaced data

The estimates of the parameters and their standard errors for 8 marginal models; 6 transitional models; and 8 random effects models for the prediction of FEV_1 are given in Tables 5.10 to 5.15. Of these 22 models given in Tables 5.10 to 5.15, 11 models had both age and years in the grain industry in the model (Tables 5.10; 5.12;

5.14) and the other 11 models (Tables 5.11; 5.13; and 5.15) were fitted when age was not included in the model. The coefficient estimates for the variable years in the grain industry for the marginal model changed when age was not in the model irrespective of the choice of covariance structure indicating that age was a potential confounder. The interaction between smoking and exposure years was significant, so it is hard to interpret the effects of main effects of exposure years and smoking status separately. A similar finding was observed for other models, i.e. the coefficient estimates for the variable years in the grain industry changed from Tables 5.12 to 5.13; and 5.14 to 5.15; so age appeared to be a potential confounder.

The concordance coefficients (r_c), for different longitudinal models of FEV_1 are given in Table 5.16. This table shows that for different covariance structures, the values of r_c ranges from 0.4554 to 0.9302. As discussed in Section 4.8.3, Chapter 4, the concordance correlation coefficient r_c behaves like R^2, the coefficient of determination in multiple regression, so one could say that the transitional models fitted for the prediction of FEV_1 gave a better fit than the other models. The values of r_c for different covariance structures for transitional models were in the range of 0.9286 to 0.9302. These values (closer to 1) showed good concordance between observed and predicted response values. On the basis of this table, one could say that the transitional models with compound symmetric covariance structure were the best for predicting the FEV_1 values of the population. However; this conclusion might not give us the true and complete picture of model building, because in longitudinal model building, it is important to test the goodness of fit of a response function and the variance covariance structure at the same time. In order to assess the adequacy of the covariance structure,

we computed the concordance coefficient $r(\hat{\omega})$ for different models, which measures the closeness between the true and assumed covariance structure. The pseudo-likelihood ratio test $\hat{\lambda}$ is used for testing the equality of true and assumed covariance structures (Table 5.17). The values of r_c (range: 0.9286 - 0.9302) and $r(\hat{\omega})$ (range: 0.9280 - 0.9361) for transitional models were very similar to each other for different covariance structures. The choice of covariance structure does not appear to have an effect on the fit of model, if previous lung function is one of the covariates in the model for analyzing the longitudinal pulmonary function data for the prediction of FEV_1.

For random effects models, these goodness of fit statistics, r_c, $r(\hat{\omega})$ and $\hat{\lambda}$ were compared with the traditional likelihood ratio test and Akaike's information criterion (Table 5.18). The difference between -2 ln (likelihood) for REM with unspecified covariance structure and REM with compound symmetric covariance structure (7908.13 - 7329.45 = 578.68) is distributed as χ^2 with 13 degrees of freedom and was highly significant at α=0.05, which indicates a REM with unspecified covariance structure gave a significantly better fit, when compared to a REM fitted with compound symmetric covariance structure. The difference between -2 ln (likelihood) for REM with autoregressive covariance structure and a REM with unspecified covariance structure (7973.42 - 7329.45 = 643.97) is distributed as χ^2 with 13 degrees of freedom and was highly significant at α=0.05, so a REM with unspecified covariance structure gave a significantly better fit, when compared to a REM fitted with autoregressive covariance structure. A similar finding was observed when a REM with an independence covariance structure was compared with a REM with an unspecified covariance structure. Among these four random effects models (with covariance

structure: independence, compound symmetric, unspecified, and autoregressive), the most parsimonious model selection was based on the maximum value of Akaike's information criterion, it was concluded that REM with unspecified covariance structure was better.

Based on r_c, $r(\hat{\omega})$ and $\hat{\lambda}$, we concluded that the random effects model with unspecified covariance structure was the best. The conclusions based on the traditional (i.e. likelihood ratio test and Akaike's Information Criterio) and new goodness of fit statistics (i.e. r_c, $r(\hat{\omega})$ and $\hat{\lambda}$) were comparable.

5.3.4.2 Marginal, transitional and random effects models for FVC assuming equally spaced data

The estimates of the parameters and their standard errors for 8 marginal; 6 transitional; and 8 random effects models for the prediction of FVC are given in Tables 5.19 to 5.24. Of these 22 models for the prediction of FVC, 11 models had both age and years in the grain industry in the model (Tables 5.19; 5.21; 5.23) and the other 11 models (Tables 5.20; 5.22; and 5.24) show results of models when age was not included. The coefficient estimates for the variable years in the grain industry for the marginal model changed when age was not in the model irrespective of the choice of covariance structure indicating that age was a potential confounder. This was also observed in the other models for FVC. The concordance coefficients (r_c), for different longitudinal models of FVC are given in Table 5.25. The values of r_c in the range of 0.8834 to 0.8874 implies good concordance between observed and predicted response values.

The concordance coefficient $r(\hat{\omega})$ for different models, and the pseudo-likelihood ratio test $\hat{\lambda}$ for testing the equality of true and assumed covariance structures are given in Table 5.26. Combining the results of Tables 5.25 and 5.26, one can conclude that the transitional model with independence covariance structure gave the better fit for the prediction of FVC. For random effects models for FVC, the goodness of fit statistics, r_c, $r(\hat{\omega})$ and $\hat{\lambda}$ were compared with the traditional likelihood ratio test and Akaike's information criterion (Table 5.28). Based on the likelihood ratio test, and Akaike's information criterion, the REM with unspecified covariance structure. gave a better fit. This decision was similar to what one would make based on new goodness of fit statistics (r_c, $r(\hat{\omega})$ and $\hat{\lambda}$).

The concordance correlation coefficient for the variance covariance structure, $r(\hat{\omega})$ ranges from 0.7978 to 0.9643 for longitudinal models of FEV_1 (Table 5.17) and 0.8225 to 0.9527 for longitudinal models of FVC (Table 5.26), which supports the observation by Vonesh et al (1996) that $r(\hat{\omega})$ has a limited range of values particularly when n is large. Table 5.17 and Table 5.26 also show values of $\hat{\lambda}$. Large values for $\hat{\lambda}$ may be related to a large sample size. In order to examine the behaviour of $r(\hat{\omega})$ and $\hat{\lambda}$ for small samples, an additional analysis was performed, which is described in Section 5.5.

Table 5.10 : Marginal model (age included) for the dependent variable FEV_1 with different covariance structures

	Independence	Compound symmetric	Unstructured	Autoregressive
	Estimate (SE-robust)	Estimate (SE-robust)	Estimate (SE-robust)	Estimate (SE-robust)
Constant	-1.010** (0.383)	-1.321*** (0.350)	-1.490*** (0.367)	-1.242*** (0.351)
Base height	0.035*** (0.002)	0.037*** (0.002)	0.039*** (0.002)	0.037*** (0.002)
Weight	0.00009 (0.001)	-0.002 (0.001)	-0.003** (0.001)	-0.001 (0.001)
Age	-0.033*** (0.002)	-0.031*** (0.002)	-0.031*** (0.002)	-0.032*** (0.002)
Yrs. in grain industry	0.011** (0.003)	0.008** (0.002)	0.009** (0.003)	0.009** (0.003)
Exsmokers[@]	0.158*** (0.044)	0.078* (0.035)	0.095** (0.034)	0.095** (0.036)
Smokers[@]	-0.002 (0.033)	0.0003 (0.029)	0.018 (0.031)	-0.005 (0.030)
Cycle II[#]	0.105*** (0.024)	0.152*** (0.023)	0.124*** (0.022)	0.142*** (0.023)
Cycle III[#]	-0.047 (0.026)	-0.017 (0.023)	-0.055* (0.022)	-0.041 (0.023)
Cycle IV[#]	0.131*** (0.027)	0.128*** (0.025)	0.103*** (0.024)	0.122*** (0.024)
Cycle V[#]	0.179*** (0.028)	0.195*** (0.026)	0.154*** (0.025)	0.164*** (0.025)
Yrs. in grain industry*exsmokers	-0.015*** (0.003)	-0.010*** (0.002)	-0.010*** (0.002)	-0.010*** (0.002)
Yrs. in grain industry*smokers	-0.014*** (0.003)	-0.009*** (0.002)	-0.009***[A] (0.002)	-0.010*** (0.002)

*: < 0.05; ** <0.01; ***: < 0.001; * ***: <0.0001; [@]: Reference category - Non-smokers

[#]: Reference category - Cycle I

Table 5.11 : Marginal model (age not in model) for dependent variable FEV_1 with different covariance structures

	Independence	Compound symmetric	Unstructured	Autoregressive
	Estimate (SE-robust)	Estimate (SE-robust)	Estimate (SE-robust)	Estimate (SE-robust)
Constant	-2.019*** (0.396)	-2.424*** (0.369)	-2.547*** (0.387)	-2.332*** (0.369)
Base height	0.038*** (0.002)	0.041*** (0.002)	0.042*** (0.002)	0.040*** (0.002)
Weight	-0.001 (0.001)	-0.003*** (0.001)	-0.004*** (0.001)	-0.003** (0.001)
Age	-	-	-	-
Yrs. In grain industry	-0.024*** (0.002)	-0.024*** (0.002)	-0.023*** (0.002)	-0.025*** (0.002)
Exsmokers[@]	0.058 (0.045)	-0.009 (0.036)	0.004*** (0.036)	0.0002 (0.038)
Smokers[@]	-0.042 (0.034)	-0.036 (0.030)	-0.023 (0.032)	-0.046 (0.031)
Cycle II[#]	0.117*** (0.025)	0.165*** (0.024)	0.134 (0.022)	0.153*** (0.024)
Cycle III[#]	-0.013 (0.027)	0.005 (0.024)	-0.034*** (0.023)	-0.019 (0.024)
Cycle IV[#]	0.168*** (0.028)	0.146*** (0.026)	0.113*** (0.025)	0.141*** (0.025)
Cycle V[#]	0.226*** (0.030)	0.214*** (0.027)	0.164*** (0.026)	0.189*** (0.027)
Yrs. in grain industry*exsmokers	-0.014*** (0.003)	-0.008*** (0.002)	-0.008*** (0.002)	-0.008*** (0.002)
Yrs. in grain industry*smokers	-0.014*** (0.003)	-0.008*** (0.002)	-0.008*** (0.002)	-0.008*** (0.002)

* : < 0.05 ; ** <0.01 ; ***: < 0.001 ; * ***: <0.0001 ; [@]: Reference category - Non-smokers ;
[#]: Reference category - Cycle I

Table 5.12 : Transitional model (age in model) for dependent variable FEV_1 with different covariance structures

	Independence	Compound symmetric	Unstructured[$]	Autoregressive
	Estimate (SE-robust)	Estimate (SE-robust)	Estimate (SE-robust)	Estimate (SE-robust)
Constant	-0.253* (0.102)	-0.215* (0.108)		-0.222* (0.104)
Previous FEV_1	0.877**** (0.008)	0.895**** (0.008)		0.884**** (0.008)
Base height	0.006**** (0.001)	0.005**** (0.001)		0.006**** (0.001)
Weight	-0.001**** (0.0003)	-0.001**** (0.00000007)		-0.001**** (0.00003)
Age	-0.006**** (0.001)	-0.006**** (0.001)		-0.006**** (0.001)
Yrs. In grain industry	0.018* (0.001)	0.002* (0.001)		0.002* (0.001)
Exsmokers[@]	-0.016 (0.01 6)	0.018 (0.017)		0.020 (0.016)
Smokers[@]	-0.076 (0.011)	-0.015 (0.012)		-0.015 (0.012)
Cycle II[#]	0.08**** (0.009)	0.070**** (0.008)		0.069**** (0.008)
Cycle III[#]	-0.201**** (0.009)	-0.199**** (0.009)		-0.195**** (0.009)
Cycle IV[#]	0.002 (0.009)	-0.002 (0.009)		-0.003 (0.009)
Cycle V[#]	-0.029** (0.010)	-0.029** (0.010)		-0.025* (0.010)
Yrs. in grain industry*exsmokers	-0.002 (0.001)	-0.002* (0.001)		-0.002* (0.001)
Yrs. in grain industry*smokers	-0.002 (0.001)	-0.002* (0.001)		-0.002* (0.001)

*: < 0.05 ; ** <0.01 ; ***: < 0.001 ; * ***: <0.0001 ; [@]: Reference category - Non-smokers

[#]: Reference category - Cycle I

$: This model was not fitted due to numerical problems.

Table 5.13 : Transitional model (age not in model) for dependent variable FEV_1 with different covariance structures

	Independence	Compound symmetric	Unstructured[$]	Autoregressive
	Estimate (SE-robust)	Estimate (SE-robust)	Estimate (SE-robust)	Estimate (SE-robust)
Constant	-0.407*** (0.102)	-0.355*** (0.107)		-0.373*** (0.104)
Previous FEV_1	0.893**** (0.007)	0.909**** (0.007)		0.899**** (0.008)
Base height	0.006**** (0.001)	0.005**** (0.001)		0.006**** (0.001)
Weight	-0.002**** (0.0003)	-0.002**** (0.00000007)		-0.002**** (0.0000)
Age	-	-	-	-
Yrs. In grain industry	-0.004**** (0.001)	-0.003**** (0.001)		-0.004**** (0.001)
Exsmokers[@]	-0.001 (0.016)	-0.00006 (0.016)		0.001 (0.016)
Smokers[@]	-0.022* (0.011)	-0.022 (0.012)		-0.022 (0.012)
Cycle II[#]	0.077**** (0.009)	0.071**** (0.008)		0.070**** (0.008)
Cycle III[#]	-0.198**** (0.009)	-0.196**** (0.009)		-0.193**** (0.009)
Cycle IV[#]	0.006 (0.009)	0.001 (0.009)		0.001 (0.009)
Cycle V[#]	-0.025* (0.010)	-0.024* (0.010)		-0.022* (0.010)
Yrs. in grain industry*exsmokers	-0.001 (0.001)	-0.002 (0.001)		-0.002 (0.001)
Yrs. in grain industry*smokers	-0.002 (0.001)	-0.002 (0.001)		-0.002 (0.001)

* ; < 0.05 ; ** <0.01 ; ***: < 0.001 ; * ***: <0.0001 ; @: Reference category - Non-smokers ; #: Reference category - Cycle I

$: This model was not fitted due to numerical problems.

Table 5.14. Random effect model (age in model) for dependent variable FEV_1 with different within-subject covariance structures and unspecified between-subject covariance structure.

	Independence	Compound symmetric	Unstructured	Autoregressive
	Estimate (SE)	Estimate (SE)	Estimate (SE)	Estimate (SE)
Constant	-1.585**** (0.265)	-1.408**** (0.272)	-1.585**** (0.265)	-1.374**** (0.271)
Base height	0.039**** (0.002)	0.038**** (0.002)	0.039**** (0.002)	0.038**** (0.002)
Weight	-0.003**** (0.0007)	-0.002* (0.001)	-0.003**** (0.001)	-0.001# (0.001)
Age	-0.031**** (0.001)	-0.030**** (0.001)	-0.031**** (0.001)	-0.031**** (0.001)
Yrs. in grain industry	0.009**** (0.002)	0.008*** (0.002)	0.009**** (0.002)	0.008 (0.002)
Exsmokers@	0.093** (0.0294)	0.072* (0.032)	0.093** (0.029)	0.082* (0.032)
Smokers@	0.014 (0.025)	-0.002 (0.026)	0.014 (0.025)	-0.006 (0.027)
Cycle II#	0.126**** (0.022)	0.156**** (0.017)	0.126**** (0.022)	0.154**** (0.016)
Cycle III#	-0.052* (0.022)	-0.013 (0.019)	-0.052* (0.022)	-0.030 (0.019)
Cycle IV#	0.103**** (0.024)	0.129**** (0.019)	0.103**** (0.024)	0.130**** (0.020)
Cycle V#	0.153**** (0.025)	0.195**** (0.020)	0.153**** (0.025)	0.166**** (0.022)
Yrs. in grain industry*exsmokers	-0.010**** (0.002)	-0.009**** (0.002)	-0.010**** (0.002)	-0.010**** (0.002)
Yrs. in grain industry*smokers	-0.008**** (0.002)	-0.008**** (0.002)	-0.008**** (0.002)	-0.007**** (0.002)

* : < 0.05 ; ** <0.01 ; ***: < 0.001 ; * ***: <0.0001 ; @: Reference category - Non-smokers ; #: Reference category - Cycle I

Table 5.15 : Random effect model (age not in model) for dependent variable FEV_1 with different within-subject covariance structures and unspecified between-subject covariance structure.

	Independence	Compound symmetric	Unstructured	Autoregressive
	Estimate (SE)	Estimate (SE)	Estimate (SE)	Estimate (SE)
Constant	-2.342****	-2.483****	-2.650****	-2.465****
	(0.259)	(0.279)	(0.272)	(0.276)
Base height	0.040****	0.041****	0.042****	0.041****
	(0.002)	(0.002)	(0.002)	(0.002)
Weight	-0.002*	-0.003****	-0.004****	-0.003****
	(0.0007)	(0.001)	(0.001)	(0.001)
Age	-	-	-	-
Yrs. in grain industry	-0.021****	-0.023****	-0.022****	-0.024****
	(0.002)	0.002	(0.002)	(0.002)
Exsmokers[@]	0.003	-0.005	0.008	-0.001
	(0.031)	(0.032)	(0.030)	(0.032)
Smokers[@]	-0.049*	-0.038	-0.027	-0.047
	(0.024)	(0.026)	(0.026)	(0.027)
Cycle II[#]	0.153****	0.163****	0.138****	0.158****
	(0.017)	(0.017)	(0.022)	(0.016)
Cycle III[#]	-0.011	0.001	-0.033	-0.018
	(0.020)	(0.019)	(0.022)	(0.019)
Cycle IV[#]	0.11****	0.134****	0.108****	0.133****
	(0.021)	(0.020)	(0.024)	(0.021)
Cycle V[#]	0.16****	0.199****	0.156****	0.172****
	(0.022)	(0.021)	(0.025)	(0.022)
Yrs. in grain industry*exsmokers	-0.007**	-0.007***	-0.007***	-0.007**
	(0.002)	(0.002)	(0.002)	(0.002)
Yrs. in grain industry*smokers	-0.005*	-0.007***	-0.007***	-0.006**
	(0.002)	(0.002)	(0.002)	(0.002)

*: < 0.05 ; ** <0.01 ; ***: < 0.001 ; * ***: <0.0001 ; [@]: Reference category - Non-smokers ;
[#]: Reference category - Cycle I

Table 5.15 : Random effect model (age not in model) for dependent variable FEV_1 with different within-subject covariance structures and unspecified between-subject covariance structure.

	Independence	Compound symmetric	Unstructured	Autoregressive
	Estimate (SE)	Estimate (SE)	Estimate (SE)	Estimate (SE)
Constant	-2.342****	-2.483****	-2.650****	-2.465****
	(0.259)	(0.279)	(0.272)	(0.276)
Base height	0.040****	0.041****	0.042****	0.041****
	(0.002)	(0.002)	(0.002)	(0.002)
Weight	-0.002*	-0.003****	-0.004****	-0.003****
	(0.0007)	(0.001)	(0.001)	(0.001)
Age	-	-	-	-
Yrs. in grain industry	-0.021****	-0.023****	-0.022****	-0.024****
	(0.002)	0.002	(0.002)	(0.002)
Exsmokers[@]	0.003	-0.005	0.008	-0.001
	(0.031)	(0.032)	(0.030)	(0.032)
Smokers[@]	-0.049*	-0.038	-0.027	-0.047
	(0.024)	(0.026)	(0.026)	(0.027)
Cycle II[#]	0.153****	0.163****	0.138****	0.158****
	(0.017)	(0.017)	(0.022)	(0.016)
Cycle III[#]	-0.011	0.001	-0.033	-0.018
	(0.020)	(0.019)	(0.022)	(0.019)
Cycle IV[#]	0.11****	0.134****	0.108****	0.133****
	(0.021)	(0.020)	(0.024)	(0.021)
Cycle V[#]	0.16****	0.199****	0.156****	0.172****
	(0.022)	(0.021)	(0.025)	(0.022)
Yrs. in grain industry*exsmokers	-0.007**	-0.007***	-0.007***	-0.007**
	(0.002)	(0.002)	(0.002)	(0.002)
Yrs. in grain industry*smokers	-0.005*	-0.007***	-0.007***	-0.006**
	(0.002)	(0.002)	(0.002)	(0.002)

*: < 0.05 ; ** <0.01 ; ***: < 0.001 ; * ***: <0.0001 ; [@]: Reference category - Non-smokers ;
[#]: Reference category - Cycle I

Table 5.16 : Concordance coefficient (r_c) for the longitudinal models of FEV_1 (with and without age).*

	Marginal		Transitional		Random effect model	
	Age in the model	Age not in the model	Age in the model	Age not in the model	Age in the model	Age not in the model
Independence	0.5766	0.5119	0.9297	0.9286	0.5433	0.4554
Compound Symmetric	0.5647	0.4880	0.9302	0.9292	0.5616	0.4788
Autoregressive equally spaced	0.5694	0.4882	0.9298	0.9288	0.5647	0.4815
Unspecified	0.5668	0.4856	-●	-●	0.5648	0.4770

* : Details of the models are shown in Tables 5.10 to 5.15

●: This model was not fitted due to numerical problems.

Table 5.17 : Concordance coefficient (r_w) for variance-covariance structure and pseudo-likelihood ratio test (λ) for longitudinal models of FEV_1(with and without age)*.

	Marginal		Transitional		Random effect model	
	Age in the model r_w (λ)	Age not in the model r_w (λ)	Age in the model r_w (λ)	Age not in the model r_w (λ)	Age in the model r_w (λ)	Age not in the model r_w (λ)
Independence	0.8003 (11263.76)	0.7978 (10275.93)	0.9355 (6887.66)	0.9361 (7332.83)	0.9249 (8300.00)	0.9314 (6282.99)
Compound Symmetric	0.9392 (3463.44)	0.9350 (3326.02)	0.9299 (7475.06)	0.9322 (7799.93)	0.9638 (4016.53)	0.9632 (3617.22)
Autoregressive equally spaced	0.9367 (3495.65)	0.9343 (3288.58)	0.9280 (6897.21)	0.9297 (7198.03)	0.9643 (3637.48)	0.9632 (3387.66)
Unspecified	0.9501 (2258.60)	0.9440 (2381.72)	-●	-●	-#	-#

* : Details of the models are shown in Tables 5.10 to 5.15

●: This model was not fitted due to numerical problems.

#: Not applicable; for REM unspecified covariance structure was the gold standard for independence, compound symmetric, and autoregressive covariance structures.

Table 5.18 : -2 ln(likelihood) and Akaikes' information criterion values for random effects models for the prediction of FEV_1

Covariance structure	-2 ln (likelihood)		Akaike's information criterion	
	Age in model	Age not in model	Age in model	Age not in model
Independence	8459.02	8695.71	-4231.51	-4349.86
Compound symmetric	7908.13	8280.88	-3957.06	-4143.44
Autoregressive equally spaced	7973.42	8360.64	-3989.71	-4183.32
Unspecified	7329.45	7752.73	-3680.73	-3892.36

Table 5.19 : Marginal model (age in model) with different covariance structures and dependent variable FVC

	Independence	Compound symmetric	Unstructured	Autoregressive
	Estimate (SE-robust)	Estimate (SE-robust)	Estimate (SE-robust)	Estimate (SE-robust)
Constant	-3.514*** (0.445)	-3.621*** (0.451)	-4.068*** (0.483)	-3.561*** (0.443)
Base height	0.054*** (0.003)	0.054*** (0.003)	0.058*** (0.003)	0.054*** (0.003)
Weight	0.001 (0.001)	0.0001 (0.001)	-0.002 (0.001)	0.001 (0.001)
Age	-0.027*** (0.002)	-0.025*** (0.002)	-0.025*** (0.002)	-0.025*** (0.002)
Yrs. In grain industry	0.014** (0.004)	0.012** (0.003)	0.012** (0.003)	0.011** (0.003)
Exsmokers[@]	0.194*** (0.054)	0.151** (0.044)	0.165*** (0.041)	0.156** (0.046)
Smokers[@]	0.031 (0.041)	0.027 (0.037)	0.047 (0.037)	0.020 (0.038)
Cycle II[#]	0.163*** (0.032)	0.259*** (0.032)	0.195*** (0.030)	0.245*** (0.032)
Cycle III[#]	-0.072* (0.034)	-0.024 (0.031)	-0.102** (0.030)	-0.077* (0.031)
Cycle IV[#]	0.119** (0.034)	0.162*** (0.033)	0.109** (0.032)	0.142*** (0.033)
Cycle V[#]	0.115** (0.036)	0.165*** (0.034)	0.093** (0.033)	0.125*** (0.034)
Yrs. in grain industry*exsmokers	-0.014*** (0.003)	-0.012*** (0.003)	-0.013*** (0.003)	-0.012*** (0.003)
Yrs. in grain industry*smokers	-0.008* (0.003)	-0.005 (0.003)	-0.006* (0.003)	-0.004 (0.003)

* : < 0.05 ; ** <0.01 ; ***: < 0.001 ; * ***: <0.0001 ; [@]: Reference category - Non-smokers

[#]: Reference category - Cycle I

Table 5.20 : Marginal model (age not in model) with different covariance structures and dependent variable FVC

	Independence	Compound symmetric	Unstructured	Autoregressive
	Estimate (SE-robust)	Estimate (SE-robust)	Estimate (SE-robust)	Estimate (SE-robust)
Constant	-4.341**** (0.454)	-4.459**** (0.461)	-4.860**** (0.497)	-4.388**** (0.453)
Base height	0.056**** (0.003)	0.057**** (0.003)	0.060**** (0.003)	0.056**** (0.003)
Weight	-0.001 (0.001)	-0.001 (0.001)	-0.003** (0.001)	-0.0007 (0.001)
Age	-	-	-	-
Yrs. in grain industry	-0.015**** (0.003)	-0.014**** (0.002)	-0.014**** (0.002)	-0.016**** (0.002)
Exsmokers@	0.111* (0.054)	0.079 (0.045)	0.091* (0.041)	0.079 (0.046)
Smokers@	-0.002 (0.041)	-0.001 (0.037)	0.014 (0.037)	-0.011 (0.038)
Cycle II#	0.172**** (0.032)	0.271**** (0.032)	0.199**** (0.030)	0.258**** (0.033)
Cycle III#	-0.045 (0.034)	-0.003 (0.032)	-0.089** (0.031)	-0.056 (0.032)
Cycle IV#	0.148**** (0.035)	0.181**** (0.033)	0.116*** (0.032)	0.162**** (0.033)
Cycle V#	0.152**** (0.037)	0.187**** (0.035)	0.099** (0.033)	0.151**** (0.034)
Yrs. in grain industry*exsmokers	-0.013*** (0.003)	-0.011**** (0.003)	-0.011**** (0.003)	-0.010*** (0.003)
Yrs. in grain industry*smokers	-0.008* (0.003)	-0.005 (0.003)	-0.005* (0.003)	-0.004 (0.003)

* : < 0.05 ; ** <0.01 ; ***: < 0.001 ; * ***: <0.0001 ; @: Reference category - Non-smokers

#: Reference category - Cycle I

Table 5.21 : Transitional model (age in model) with different covariance structures and dependent variable FVC

	Independence	Compound symmetric	Unstructured[#]	Autoregressive
	Estimate (SE-robust)	Estimate (SE-robust)	Estimate (SE-robust)	Estimate (SE-robust)
Constant	-1.188**** (0.153)	-1.120**** (0.163)		-1.183**** (0.167)
Previous FVC	0.797**** (0.009)	0.817**** (0.008)		0.806**** (0.008)
Base height	0.015**** (0.001)	0.014**** (0.001)		0.015**** (0.001)
Weight	-0.002**** (0.0004)	-0.002**** (0.0004)		-0.002**** (0.0004)
Age	-0.008**** (0.001)	-0.008**** (0.001)		-0.009**** (0.001)
Yrs. In grain industry	0.002 (0.001)	0.002 (0.001)		0.003** (0.001)
Exsmokers[@]	0.031**** (0.023)	0.028 (0.025)		0.030 (0.024)
Smokers[@]	-0.005**** (0.015)	-0.006 (0.017)		0.001 (0.017)
Cycle II[#]	0.117**** (0.014)	0.102**** (0.013)		0.090**** (0.013)
Cycle III[#]	-0.268**** (0.014)	-0.271**** (0.014)		-0.268**** (0.014)
Cycle IV[#]	0.050 (0.013)	0.036** (0.013)		0.031* (0.013)
Cycle V[#]	-0.035 (0.014)	-0.039** (0.013)		-0.044**** (0.013)
Yrs. in grain industry*exsmokers	-0.002 (0.001)	-0.002 (0.001)		-0.002 (0.001)
Yrs. in grain industry*smokers	0.0001 (0.001)	0.00003 (0.001)		-0.0005 (0.001)

* : <0.05; **: < 0.01; *** : < 0.001 ; ****: <0.0001 ; [@]: Reference category - Non-smokers ; [#]: Reference category -Cycle I

$: This model was not fitted due to numerical problems.

Table 5.22 : Transitional model (age not in model) with different covariance structures and dependent variable FVC

	Independence	Compound symmetric	Unstructured$	Autoregressive
	Estimate (SE-robust)	Estimate (SE-robust)	Estimate (SE-robust)	Estimate (SE-robust)
Constant	-1.404**** (0.156)	-1.333**** (0.163)		-1.399**** (0.168)
Previous FVC	0.808**** (0.009)	0.827**** (0.008)		0.817**** (0.008)
Base height	0.015**** (0.001)	0.015**** (0.001)		0.015**** (0.001)
Weight	-0.002**** (0.0004)	-0.003**** (0.0004)		-0.003**** (0.0004)
Age	-	-		-
Yrs. In grain industry	-0.007**** (0.001)	-0.007**** (0.001)		-0.006**** (0.001)
Exsmokers@	0.004 (0.023)	0.001 (0.024)		0.002 (0.024)
Smokers@	-0.015 (0.015)	-0.017 (0.016)		-0.011 (0.017)
Cycle II#	0.119**** (0.014)	0.106**** (0.013)		0.094**** (0.013)
Cycle III#	-0.263**** (0.014)	-0.264**** (0.014)		-0.261**** (0.014)
Cycle IV#	0.057**** (0.013)	0.046*** (0.013)		0.040 (0.013)
Cycle V#	-0.026 (0.014)	-0.028* (0.013)		-0.033* (0.013)
Yrs. in grain industry*exsmokers	-0.002 (0.001)	-0.001 (0.001)		-0.002 (0.001)
Yrs. in grain industry*smokers	0.0001 (0.001)	0.0001 (0.001)		-0.0003 (0.001)

*: < 0.05 ; **: < 0.01 ; ***: < 0.001 ; ****: <0.0001 ; @: Reference category - Non-smokers ;

#: Reference category - Cycle I ;

$: This model was not fitted due to numerical problems.

Table 5.23 : Random effect model (age in model) with independence covariance structure for between-subject and dependent variable FVC

	Independence	Compound symmetric	Unstructured	Autoregressive
	Estimate (SE)	Estimate (SE)	Estimate (SE)	Estimate (SE)
Constant	-3.669**** (0.313)	-3.682**** (0.336)	-4.118**** (0.321)	-3.629**** (0.331)
Base height	0.053**** (0.002)	0.055**** (0.002)	0.058**** (0.002)	0.054**** (0.002)
Weight	0.002** (0.001)	-0.0002 (0.001)	-0.002* (0.001)	0.0003 (0.001)
Age	-0.021**** (0.002)	-0.024**** (0.002)	-0.025**** (0.002)	-0.024**** (0.002)
Yrs. in grain industry	0.012*** (0.003)	0.012**** (0.003)	0.012**** (0.003)	0.011*** (0.003)
Exsmokers@	0.134** (0.40)	0.142** (0.041)	0.165**** (0.038)	0.144*** (0.042)
Smokers@	0.009 (0.030)	0.029 (0.033)	0.045 (0.032)	0.020 (0.035)
Cycle II#	0.270**** (0.022)	0.291**** (0.022)	0.198**** (0.030)	0.289**** (0.022)
Cycle III#	-0.016 (0.026)	0.003 (0.025)	-0.099** (0.030)	-0.039 (0.025)
Cycle IV#	0.154**** (0.027)	0.187**** (0.026)	0.111*** (0.032)	0.178**** (0.027)
Cycle V#	0.149**** (0.028)	0.189**** (0.027)	0.093** (0.032)	0.150**** (0.028)
Yrs. in grain industry*exsmokers	-0.013**** (0.003)	-0.012**** (0.003)	-0.013**** (0.002)	-0.011**** (0.003)
Yrs. in grain industry*smokers	-0.004 (0.003)	-0.005 (0.003)	-0.006* (0.002)	-0.004 (0.003)

* : < 0.05 ; ** : < 0.01 ; *** : < 0.001 ; ****: <0.0001 ; @: Reference category - Non-smokers ;
#: Reference category - Cycle I

Table 5.24 : Random effect model (age not in model) with independence covariance structure for between-subject and dependent variable FVC

	Independence	Compound symmetric	Unstructured	Autoregressive
	Estimate (SE)	Estimate (SE)	Estimate (SE)	Estimate (SE)
Constant	-4.257**** (0.312)	-4.504**** (0.336)	-4.920**** (0.323)	-4.448**** (0.330)
Base height	0.054**** (0.002)	0.057**** (0.002)	0.060**** (0.002)	0.056**** (0.002)
Weight	0.001 (0.001)	-0.001 (0.001)	-0.002* (0.001)	-0.0006 (0.001)
Age	-	-	-	-
Yrs. in grain industry	-0.009**** (0.002)	-0.012**** (0.002)	-0.013**** (0.002)	-0.014**** (0.002)
Exsmokers@	0.088* (0.040)	0.080 (0.041)	0.098* (0.038)	0.080 (0.042)
Smokers@	-0.008 (0.030)	0.001 (0.034)	0.013 (0.032)	-0.010 (0.035)
Cycle II#	0.277**** (0.022)	0.300**** (0.022)	0.205**** (0.030)	0.295**** (0.022)
Cycle III#	-0.004 (0.026)	0.019 (0.025)	-0.085** (0.030)	-0.025 (0.025)
Cycle IV#	0.159**** (0.027)	0.199**** (0.026)	0.117**** (0.032)	0.188**** (0.027)
Cycle V#	0.154**** (0.028)	0.201**** (0.027)	0.098** (0.033)	0.164**** (0.029)
Yrs. in grain industry*exsmokers	-0.012**** (0.003)	-0.011**** (0.003)	-0.012**** (0.002)	-0.010**** (0.003)
Yrs. in grain industry*smokers	-0.004 (0.003)	-0.005 (0.003)	-0.005* (0.002)	-0.004 (0.003)

* : < 0.05 ; ** : < 0.01 ; *** : < 0.001 ; ****: <0.0001 ; @: Reference category - Non-smokers ;
#: Reference category - Cycle I

Table 5.25 : Concordance coefficient (r_c) for the longitudinal models of FVC with or without age. *

	Marginal		Transitional		Random effect model	
	Age in the model	Age not in the model	Age in the model	Age not in the model	Age in the model	Age not in the model
Independence	0.4683	0.4300	0.8869	0.8851	0.4398	0.3972
Compound Symmetric	0.4633	0.4212	0.8834	0.8865	0.4584	0.4168
Autoregressive equally spaced	0.4651	0.4257	0.8874	0.8857	0.4676	0.4199
Unspecified	0.4710	0.4269	-[e]	-[e]	0.4704	0.4254

* : Details of the models are shown in Tables 5.18 to 5.23

[e]: Numerical problems to fit this model, convergence criterion not met

Table 5.26 : Concordance coefficient (r_w) for variance-covariance structure and pseudo-likelihood ratio test (λ), for longitudinal models of FVC with or without age. *

	Marginal		Transitional		Random effect model	
	Age in the model r_w (λ)	Age not in the model r_w (λ)	Age in the model r_w (λ)	Age not in the model r_w (λ)	Age in the model r_w (λ)	Age not in the model r_w (λ)
Independence	0.8268 (9929.95)	0.8225 (9161.58)	0.9328 (5008.50)	0.9334 (4982.80)	0.9105 (9577.99)	0.9129 (7896.80)
Compound Symmetric	0.9206 (4282.26)	0.9146 (4121.64)	0.9277 (5656.88)	0.9280 (5599.02)	0.9527 (4743.3863)	0.9515 (4385.37)
Autoregressive equally spaced	0.9169 (4516.85)	0.9119 (4269.76)	0.9156 (5981.89)	0.9163 (5938.71)	0.9496 (5097.36)	0.9482 (4766.98)
Unspecified	0.9333 (2577.66)	0.9231 (2681.21)	-$	-$	-#	-#

* : Details of the models are shown in Tables 5.18 - 5.23

$: This model was not fitted due to numerical problems.

#: For REM unspecified covariance structure was the gold standard.

Table 5.27 : -2 ln(likelihood) and Akaikes' information criterion values for random effects models for the prediction of FVC

Covariance structure	-2 ln (likelihood)		Akaike's information criterion	
	Age in model	Age not in model	Age in model	Age not in model
Independence	11365.62	11447.60	-5684.81	-5725.80
Compound symmetric	10871.29	11017.24	-5438.64	-5511.62
Autoregressive equally spaced	11023.13	11170.96	-5514.57	-5588.48
Unspecified	10156.80	10336.77	-5094.40	-5184.38

5.4. Random Effects Model for Unequally Spaced Data.

The results of random effects models assuming unequally-spaced AR(1) and unequally-spaced AR(1) with observational error are given in Tables 5.28 to 5.34. Tables 5.28 to 5.31 contain the values of parameter estimates and their standard errors obtained by fitting the random effects model for the prediction of FEV_1 and FVC, with (Tables 5.28 and 5.30) and without (Table 5.29 and 5.31) age in the model. Model concordance coefficients (r_c); variance-covariance concordance coefficients $r(\hat{\omega})$ for measuring the closeness and pseudo likelihood ratio test $\hat{\lambda}$ for testing the equality of two covariance structures are given in Table 5.32. For FEV_1 and FVC, the values of $r(\hat{\omega})$ are closer to 1. The values of $\hat{\lambda}$ for FEV_1 show that these two covariance structures were not significantly different, but this was not true for FVC. The values of $\hat{\lambda}$ for FVC indicate that these two covariance structures were statistically different. Based on the values of r_c, $r(\hat{\omega})$, and $\hat{\lambda}$, we concluded that for the prediction of FEV_1, the unequally-spaced AR(1) covariance structure is adequate to fit the random effect model with age as a random effect. For the prediction of FVC, unequally-spaced AR(1) with observational error covariance structure gave a better fit. The values of $\hat{\lambda}$ for the random effects model for FVC were comparatively higher than what we obtained for the random effects model for FEV_1. One possible reason for these high values could be that measurement of FVC may be associated with greater measurement/observational error than those observed for FEV_1. This in fact was supported by the estimate of observational error for FVC (=0.3870), which was higher than the estimate

Table 5.28 : Random effect model (age in model) with unspecified covariance structure for between-subject and dependent variable FEV_1

	Unequally-spaced Autoregressive		Unequally-spaced Autoregressive with observational error	
	Estimate	S.E.	Estimate	S.E.
Constant	- 1.540****	0.255	- 1.540****	0.255
Base height	0.0378****	0.002	0.038****	0.002
Weight	- 0.001	0.001	- 0.001	0.001
Age	- 0.028****	0.002	- 0.028****	0.002
Yrs. in grain industry	0.006*	0.003	0.006*	0.003
Exsmokers[@]	0.066*	0.031	0.065*	0.031
Smokers[@]	- 0.026	0.023	- 0.026	0.023
Cycle II[#]	0.149****	0.017	0.149****	0.017
Cycle III[#]	- 0.020	0.019	- 0.019	0.019
Cycle IV[#]	0.112****	0.020	0.111****	0.020
Cycle V[#]	0.171****	0.022	0.170****	0.022
Yrs. in grain industry*exsmokers	- 0.009****	0.002	- 0.009****	0.002
Yrs. in grain industry*smokers	- 0.005*	0.002	- 0.005*	0.002

* : < 0.05 ; ** : < 0.01 ; *** : < 0.001 ; ****: <0.0001 ; [@]: Reference category - Non-smokers ;
[#]: Reference category - Cycle I

Table 5.29 : Random effect model (age not in model) with unspecified covariance structure for between-subject and dependent variable FEV_1

	Unequally-spaced Autoregressive		Unequally-spaced Autoregressive with observational error	
	Estimate	S.E.	Estimate	S.E.
Constant	-2.343****	0.259	-2.344****	0.259
Base height	0.040****	0.002	0.040****	0.002
Weight	-0.002*	0.001	-0.002*	0.001
Age	-	-	-	-
Yrs. in grain industry	-0.021****	0.002	-0.021****	0.002
Exsmokers@	0.004	0.031	0.003	0.031
Smokers@	-0.049*	0.024	-0.049*	0.024
Cycle II#	0.153****	0.017	0.153****	0.017
Cycle III#	-0.011	0.020	-0.010	0.020
Cycle IV#	0.110****	0.021	0.109****	0.021
Cycle V#	0.166****	0.022	0.0166****	0.022
Yrs. in grain industry*exsmokers	-0.007**	0.002	-0.007**	0.002
Yrs. in grain industry*smokers	-0.005**	0.002	-0.005**	0.002

* : < 0.05 ; ** : < 0.01 ; *** : < 0.001 ; ****: <0.0001 ; @: Reference category - Non-smokers ;
#: Reference category - Cycle I

Table 5.30 : Random effect model (age in model) with unspecified covariance structure for between-subject and dependent variable FVC

	Unequally-spaced Autoregressive		Unequally-spaced Autoregressive with observational error	
	Estimate	S.E.	Estimate	S.E.
Constant	-3.675****	0.313	-3.664****	0.336
Base height	0.0530****	0.002	0.054****	0.002
Weight	0.002**	0.001	-0.0001	0.001
Age	-0.021****	0.002	-0.024****	0.002
Yrs. in grain industry	0.012***	0.003	0.012****	0.003
Exsmokers[@]	0.0140***	0.040	0.147***	0.042
Smokers[@]	0.009****	0.030	0.028	0.034
Cycle II[#]	0.269****	0.022	0.294****	0.022
Cycle III[#]	-0.012	0.026	0.001	0.025
Cycle IV[#]	0.155****	0.027	0.191****	0.026
Cycle V[#]	0.149****	0.028	0.186****	0.027
Yrs. in grain industry*exsmokers	-0.013****	0.003	-0.012****	0.003
Yrs. in grain industry*smokers	-0.004	0.003	-0.005	0.003

*: < 0.05 ; **: < 0.01 ; ***: < 0.001 ; ****: <0.0001 ; [@]: Reference category - Non-smokers ;
[#]: Reference category - Cycle I

Table 5.31 : Random effect model (age not in model) with unspecified covariance structure for between-subject and dependent variable FVC

	Unequally-spaced Autoregressive		Unequally-spaced Autoregressive with observational error	
	Estimate	S.E.	Estimate	S.E.
Constant	-4.264****	0.312	-4.497****	0.336
Base height	0.054****	0.002	0.057****	0.002
Weight	0.001	0.001	-0.001	0.001
Age	-	-	-	-
Yrs. in grain industry	-0.009***	0.002	-0.013****	0.002
Exsmokers[@]	0.093*	0.040	0.083*	0.042
Smokers[@]	-0.008	0.030	-0.0005	0.034
Cycle II[#]	0.277****	0.022	0.303****	0.022
Cycle III[#]	-0.003	0.026	0.017	0.025
Cycle IV[#]	0.160****	0.027	0.202****	0.026
Cycle V[#]	0.154****	0.028	0.199****	0.027
Yrs. in grain industry*exsmokers	-0.012****	0.003	-0.011****	0.003
Yrs. in grain industry*smokers	-0.004	0.003	-0.005	0.003

* : < 0.05 ; ** : < 0.01 ; *** : < 0.001 ; ****: <0.0001 ; [@]: Reference category - Non-smokers ;
[#]: Reference category - Cycle I

Table 5.32 : Model concordance coefficient (r_c) , variance-covariance concordance coefficient $r(\hat{\omega})$ and pseudo likelihood ratio test $\hat{\lambda}$ for the longitudinal models of FEV_1 and FVC with or without age.

	Unequally-spaced Autoregressive	Unequally-spaced Autoregressive with observational error	variance-covariance concordance coefficient	pseudo likelihood ratio test
	rc	rc	$r(\hat{\omega})$	$(\hat{\lambda})$
FEV_1				
Age in model	0.5435	0.5432	0.9999*	0.8053
Age not in model	0.4553	0.4550	0.9999*	1.0975
FVC				
Age in model	0.4403	0.4592	0.9551*	3737.94
Age not in model	0.4057	0.4175	0.9666*	2515.85

Table 5.33. -2 ln(likelihood) and Akaikes' information criterion values for different models.

	-2 ln (likelihood)		Akaike's information criterion	
	Unequally-spaced autoregressive	Unequally-spaced autoregressive with observational error	Unequally-spaced autoregressive	Unequally-spaced autoregressive with observational error
FEV_1 :				
Age in model	8456.24	8455.65	-4231.12	-4231.82
Age not in model	8692.89	8691.98	-4349.45	-4349.99
FVC:				
Age in model	11355.64	10850.03	-5680.82	-5429.02
Age not in model	11438.78	10998.80	-5722.39	-5503.40

observational error for FEV_1 (=0.0822). Using the likelihood ratio test and Akaike's information criterion (Table 5.34), to test for the presence of observational error, the results were similar to the above. For the prediction of FEV_1, the difference (8456.24 - 8455.65= 0.59) (Table 5.33) between -2 ln (likelihood) for unequally-spaced AR(1) and unequally-spaced AR(1) with observational error covariance structures is distributed as χ^2 with one degree of freedom and is not statistically significant. Based on the likelihood ratio test and maximum value of AIC value, a REM with unequally-spaced AR(1) covariance structure was the most appropriate. Similar conclusions were made when age was not included in the fixed part of the model, but was considered as a random effect.

For the prediction of FVC, the test for the presence of observational error involved the same steps as above. The difference (11355.64 - 10850.03 = 505.61) (Table 5.33) between -2 ln (likelihood) for unequally-spaced AR(1) and -2 ln (likelihood) for unequally-spaced AR(1) with observational error, is distributed as χ^2 with one degree of freedom, which is significant and we conclude that observational error was present. Hence for the prediction of FVC, we need to fit a REM with continuous-time AR(1) with observational error. When age was not in the fixed part of the REM (Table 5.34), similar conclusions were made about the covariance structure.

5.5 Exposure to grain dust and lung dysfunction

Among all the models, the transitional model with independence covariance structure was used to predict the annual decline in FEV_1. Age and years in the industry were highly correlated. In analyzing these data, our main interest was to study the effect

of grain dust exposure on the respiratory health of grain workers. To avoid the potential effect of the collinear variable age, the coefficients from the transitional model (age not in model - Table 5.12) were used to study the longitudinal declines in FEV_1. The predicted annual reduction in FEV_1 values by smoking status and by years in the industry were computed using the transitional model with independence covariance structure given in Table 5.12. These yearly losses in FEV_1 by smoking status and years in the grain industry are shown in Figure 5.1. Dosman and McDuffie (1987) reported that 70% of elevators had achieved dust control by 1986. Based on this information, annual declines in FEV_1 were calculated before dust control (i.e based on first three cycles; Cycle three ended on Sept 30, 1987) and after dust control (i.e based on Cycle IV and Cycle V). The declines are shown before dust control and after dust control (Figure 5.1). The graphs before dust control indicate that i) for the subjects who were in the industry for less than 10 years, the annual losses in FEV_1 was greatest in ex-smokers; ii) the non-smokers and current smokers who were in the industry for 10 to 20 years had greater annual loss in FEV_1 than did the ex-smokers; iii) in workers with 20 yr or more in the industry, the yearly decline in lung function test values were similar in ex-smokers and current smokers and higher than the yearly decline in non-smokers. The graphs after dust control indicate that there was improvement in the lung function of grain workers in all smoking and exposure categories.

Figure 5.2 shows the mean annual loss in FEV_1 by age. The yearly loss in FEV_1 increases with age. Before dust control, the yearly loss was 30 ml for grain workers who were younger than 24. The yearly loss was highest in the age group 35-39 and after that it levels off. Lung function values improved after dust control and the yearly loss in

FEV_1 was lower for all age groups when compared with yearly losses before dust control.

The transitional model with independence covariance structure was used to predict annual decline in FVC. Predicted annual reductions in FVC values stratified by smoking status and years in the industry were computed using the transitional model with independence covariance structure given in Table 5.22. These yearly losses by smoking status and years in the grain industry are shown in Figure 5.3. The declines are shown before and after dust control. The graphs before dust control indicate that i) for the subjects who were in the industry for less than 10 years, the annual loss in FVC was very similar in all three smoking groups (-26.09 ml; -25.30; and -24.19 for non-smokers, ex-smokers and current smokers respectively); ii) the ex-smokers and current smokers who were in the industry for 10 to 20 years had greater annual loss in FVC than did the non-smokers; iii) for workers with 20 yr or more in the industry, yearly declines in FVC test value were highest for ex-smokers. The graphs after dust control indicate that yearly losses in FVC were much smaller than yearly losses in FVC before dust control.

Figure 5.4 shows the mean annual loss in FVC by age. The yearly loss in FVC increases with age. Before dust control, the yearly loss was 18 ml for grain workers who were younger than 24. The yearly loss was highest in the age group 35-39 (100 ml) and after age 50 the yearly loss in FVC levels off. Lung function improved after dust control. The yearly loss in FVC was lower in all the age groups when compared with losses before dust control.

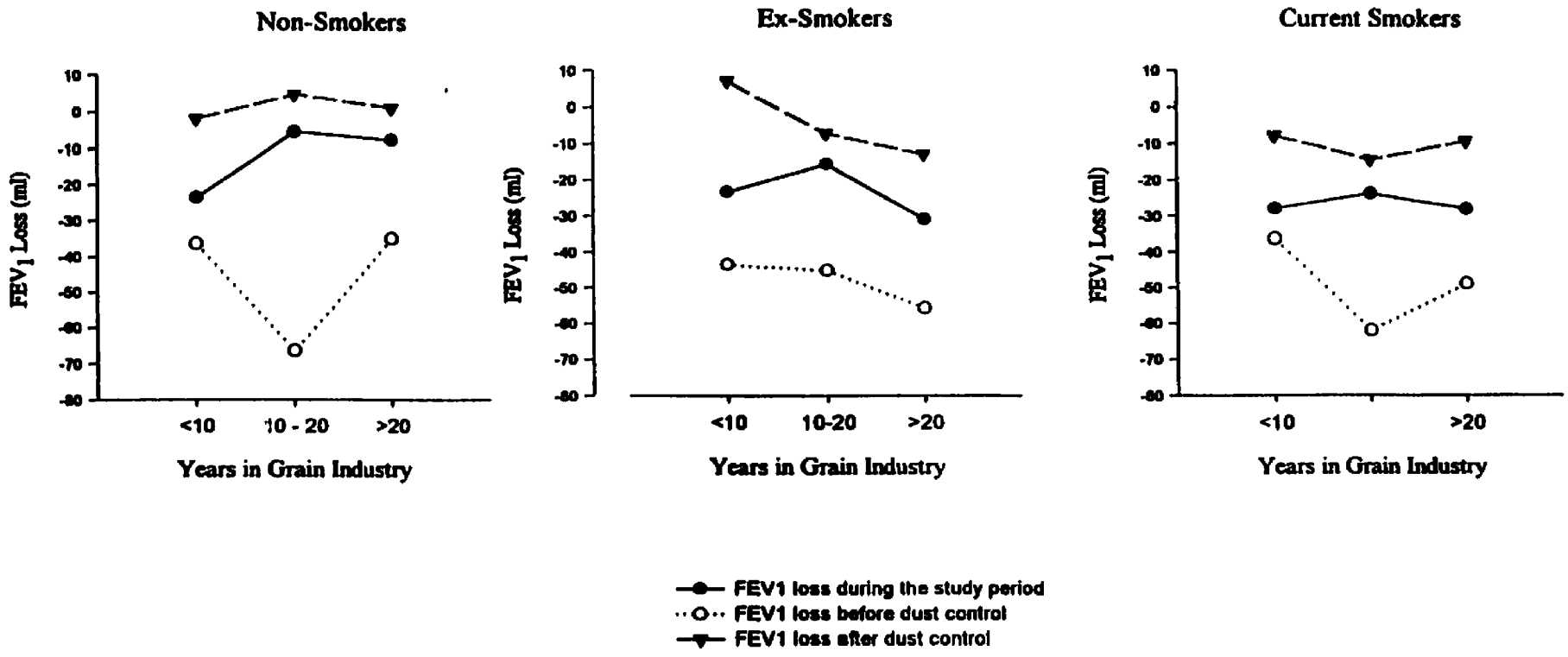


Fig. 5.1. Annual loss in FEV_1 by smoking status and years in grain industry

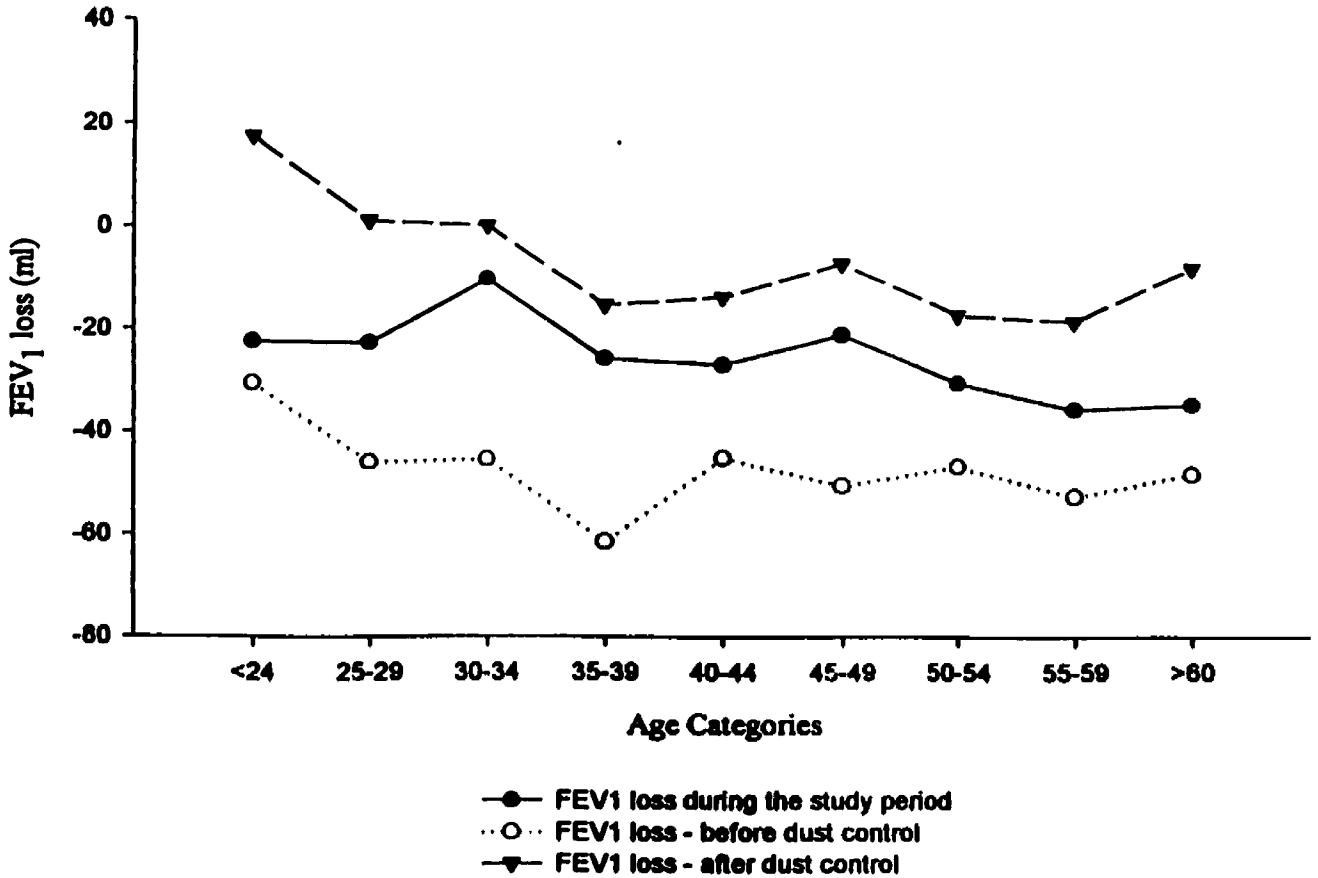


Fig. 5.2. Annual loss in FEV1 by age categories

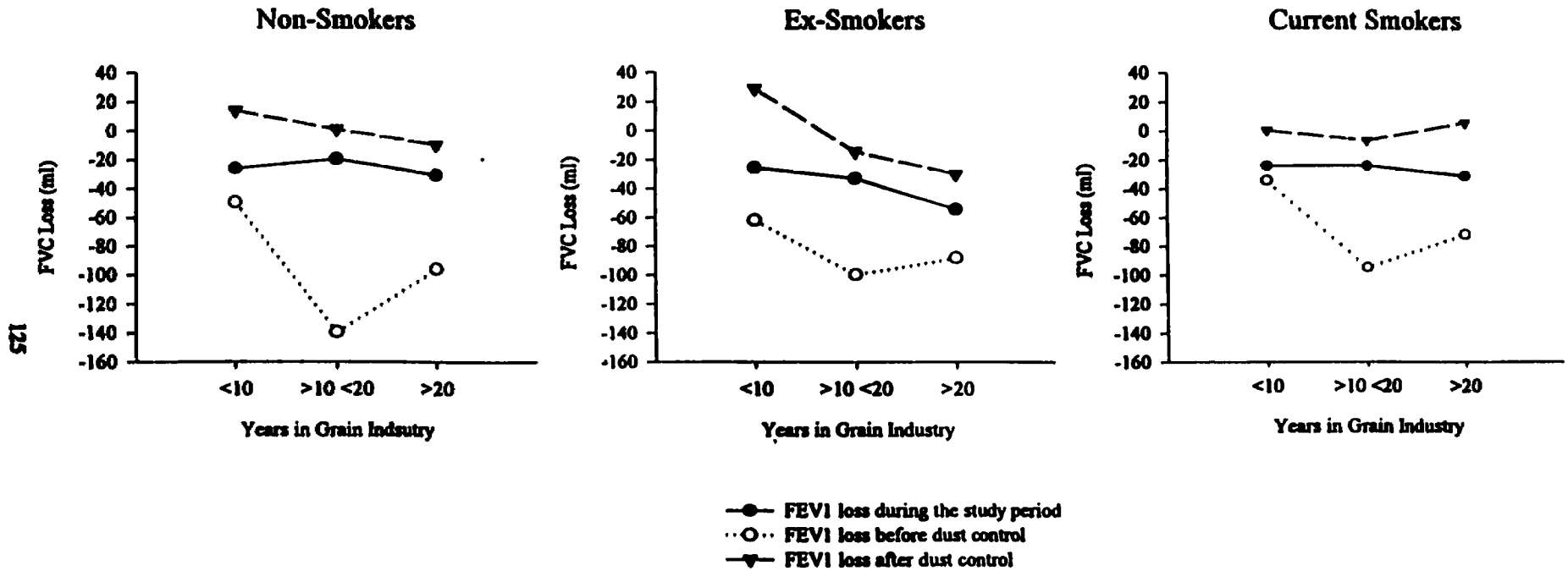


Fig. 5.3. Annual loss in FVC by smoking status and years in grain industry

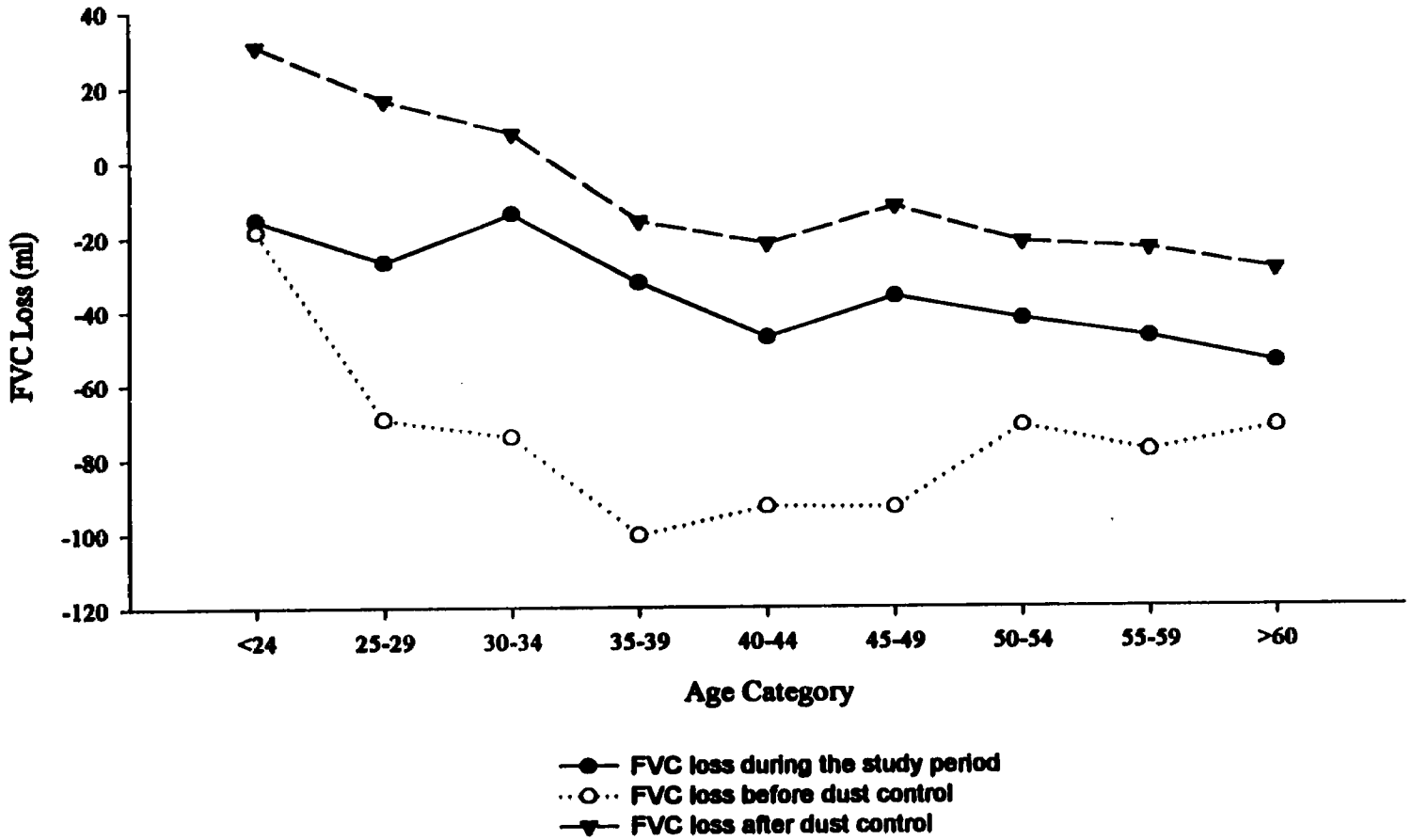


Fig. 5.4. Annual FVC loss by different age categories

6. DISCUSSION

In Chapter 5, we evaluated the relationship between lung function and years worked in the grain industry by utilizing three choices of statistical models available for the analysis of longitudinal data.

Our analysis showed that previous lung function as one of the covariates in the transitional model played an important role. One form of transitional model, also known as an autoregressive model has been used by several authors (Rosner et al., 1985; Rosner and Munoz, 1988; Ware et al., 1989; and Pahwa et al., 1994) to analyze longitudinal lung function data. These type of models were frequently used in econometrics, but Rosner and Munoz (1985) were the first to use these models in epidemiological studies to analyze equally spaced longitudinal lung function data. Rosner and Munoz (1985) used two consecutive observations available on study subjects and used the least squares approach to analyze the data. Rosner and Munoz (1988) extended the use of these statistical models for unequally spaced data. Ware et al. (1989) used these models to analyze longitudinal lung function data from Six Cities study. Pahwa et al. (1994) used first order autoregressive model to analyze longitudinal lung function data from the first three cycles of GDMSP obtained from all the provinces and territories which participated in this program. As there was no control group for grain workers, Pahwa et al. (1994) compared the yearly losses in FEV_1 in the grain workers with yearly losses in non-smoking men in Six Cities in the United States (Ware

et al., 1990). The mean FEV_1 values were higher in the younger age groups of grain workers, however after a certain age, these values were smaller when compared with the general male population of six cities. The predicted annual loss of FEV_1 was reported to be higher in the grain workers for all age groups, indicating that grain dust has an adverse health effect on the respiratory health of grain workers. Pahwa et al. (1994) showed that there were regional differences in the effects of grain dust on workers. Based on the first three observations, Dosman and McDuffie (1987) reviewed the results of Labour Canada's environmental and medical surveillance program for elevator workers. Keeping in view that there were some limitations associated with the data, they recommended that an exposure level of > 5 mg/m^3 could have negative respiratory health effects. McDuffie et al. (1992) reported the respiratory health of Canadian grain elevator workers studied on two consecutive occasions (each occasion labelled as 'cycle') by the Labour Canada Medical Surveillance program. It was reported that the prevalence of chronic cough and chronic dyspnea did not change significantly, but there was a significant increase in chronic sputum production and chronic wheeze from cycle I to cycle II. In each cycle, the obstructive dysfunction was more prevalent than the restrictive dysfunction. Huy et al. (1991) studied the association between grain dust exposure and change in lung function values. They reported that annual decline in FEV_1 for the workers in the high exposure group (mean exposure > 9 mg/m^3); the intermediate exposure group (4 to 9 mg/m^3); and the low exposure group were 34.1 ml/yr, 21.1 ml/yr and 10.4 ml/yr respectively. Huy et al. (1991) concluded that dust levels should be kept below 4 mg/m^3 to minimize chronic health effects. Some of the recommendations for reducing the effect of grain dust on the lungs were given by

Beclake et al. (1994): (i) Labour Canada and the grain industry should review the current Canadian standards for the grain dust exposure in the workplace; (ii) the permissible exposure limit (PEL) should be lowered to 5 mg/m^3 to control the short term health effects of grain dust; (iii) the Labour Canada GDMSP should be continued and regional differences in the effects of grain dust on workers should be studied and exposure-response relation should be established. A cross-sectional analysis of the dust and spirometric data showed an inverse dose-response relationship (Dosman and McDuffie, 1987). These relationships should be investigated through a longitudinal approach. One major limitation of GDMSP was that dust concentrations in the grain workers at the work place available over several time points could not be matched to individual workers in order to investigate dose-response relationships between dust and spirometric data through a longitudinal approach.

Based on the report of Dosman and McDuffie (1987), dust control was achieved in 70% of the grain elevators by implementing dust collecting, ventilation, mechanical controls and other operational procedures by the year 1986. Based on this information, Cycle III (which ended in September, 1987) was used as a cut-off point to determine whether there was any improvement in lung function (FEV_1 and FVC) after dust control (Cycle IV and Cycle V). In this thesis, only data from the province of Saskatchewan were used. After the dust was controlled in grain elevators, improvement in the respiratory health of grain workers was noticed (Figures 5.1 to 5.3, Chapter 5). The slight improvements in the mean values of FEV_1 (Fig. 5.1, Chapter 5) and FVC (Fig. 5.2) from cycle I to cycle II could be due to the learning effect. It would have been useful to test whether this learning effect had any effect on the regression coefficients.

Glindmeyer et al. (1982); and Nakadate and Kagawa (1991) did not find that the learning effect had biased the estimates of the regression parameters. Glindmeyer et al. (1982) compared the cross-sectional results at each visit with longitudinal changes by analyzing the same data. They found that the cross-sectional results were different from the longitudinal results. This difference remained even when the first visit was removed to reduce the learning effect on longitudinal estimates. Fig. 5.3 shows that after dust control was implemented (i.e after cycle III) there were there were significant reductions in the prevalence of respiratory symptoms.

The yearly loss in lung function test variables by age group and by smoking status and years in industry are shown before (i.e based on the first three cycles) and after (i.e based on cycle IV and cycle V) dust control. It was found that before dust control the yearly loss in the lung function test variables (FEV_1 and FVC) increased with increasing age. Levelling off of loss of FVC after age 50 and loss of FEV_1 after age 39 could be due to the healthy workers' effect or to job seniority which tends to lesser exposure to dust. The healthy workers' effect was shown in the population of new grain workers (Zejda et al, 1991). The levelling off of yearly losses in lung function test variables after a certain number of years in the grain industry were also reported by Pahwa et al. (1992). After the dust control measures were introduced, the yearly losses in FEV1 and FVC (Figures 5.4 and 5.5) reduced markedly in all the age groups showing that grain dust definitely had an adverse effect on the respiratory health of grain workers.

In Chapter 5, for random effects model, only the estimates of fixed effects parameters were reported, because surveillance program was conducted to study the

long term effects of grain dust on the respiratory health of grain workers population. However, the estimate of random effect parameter (i.e effect of age) can be used to study the decline in the individual grain elevator worker. In fact, if the data are unequally spaced, proper analysis for analyzing unequally-spaced data must be used. In such cases, it may be important to test for the presence of observational error. Based on GDMSP data analysis, it was concluded that inclusion of observational error in the unequally-spaced covariance structure was important while fitting the random effects model for the prediction of FVC.

The goodness of fit statistics (r_c, and $r(\phi)$) introduced by Vonesh et al. (1996) were used to assess the goodness of fit of a model, and of variance-covariance structure respectively. The null hypothesis that specified covariance structure was equal to the true covariance structure was tested using the pseudo-likelihood ratio test. The values of $r(\phi)$ had very limited range, and values of $\hat{\lambda}$ were very large, which could be due to large sample size, because $\hat{\lambda}$ is a function of sample size. In order to examine the behaviour of the values pseudo like ratio test with small samples and discrete data, additional analyses were performed on small and discrete data sets, which is discussed in the next section.

6.1 Examining the behaviour of $r(\phi)$ and $\hat{\lambda}$ with small samples and discrete data sets

Three different data sets were considered for this purpose. These data sets were: (1) data from New Grain Workers' study, explained in Section 3.3, Chapter 3; (2) data from the Carpel Tunnel Syndrome (CTS) study (Spooner et al, 1993). In the latter

study subjects were recruited over 1 year from patients who regularly received their primary health care at the ambulatory clinic of the Department of Family Medicine, University of Saskatchewan, or from patients who had been referred to the Division of Neurology, University of Saskatchewan, by their family physicians in Saskatoon for nerve conduction studies as part of a workshop for the diagnosis of CTS (Spooner et al, 1993) and (3) data from the Pothoff and Roy study (1964), which recorded the distance from the center of the pituitary to the pterygomaxillary fissure at the age of 8, 10,12 and 14 years.

For the above mentioned three data sets, marginal models were fitted assuming four different covariance structures (Uncorrelated, compound symmetric, unstructured and autoregressive) using GEE approach. Models fitted to these data are given in Table 6.1. The values of the model concordance coefficient (r_c); variance-covariance concordance coefficient ($r(\hat{\omega})$); and the pseudo-likelihood ratio test ($r(\hat{\omega})$) are given in Table 6.2. The range of $r(\hat{\omega})$ is from 0.5884 to 0.8525; 0.7161 to 0.9939; and 0.6284 to 0.9878 for CTS study; NGW study and Jaw growth study respectively. These ranges were wider than those reported for GDMSP data set indicating that the values $r(\hat{\omega})$ are influenced by sample size. The values of $\hat{\lambda}$ also varies depending on the covariance structure. The pseudo-likelihood ratio test $\hat{\lambda}$ ranges: (i) from 93.1603 to 180.5732 for CTS study; (ii) from 14.7411 to 790.2394 for NGW study and (iii) from 1.5212 to 74.7752 for jaw growth study. These results show that the values of the pseudo-likelihood ratio test $\hat{\lambda}$ do depend on the sample size and the covariance structure.

Table 6.1.: Models fitted in three different studies

	Model
	$(Difficulty\ with\ coordination)_{ij} = \beta_0 + \beta_1\ (treatment)_{ij} + \beta_2\ (MDMLAL)_{ij} + \beta_3\ (MDMLAR)_{ij} + \beta_4\ (MDMLBL)_{ij} + \beta_5\ (MDMLBR)_{ij} + \beta_6\ (MDPLAL)_{ij} + \beta_7\ (MDPLBR)_{ij} + \beta_8\ (HANDLR)_{ij} + \varepsilon_{ij}$
Carpel Tunnel Syndrome Study (n = 96)	where MDMLAL : MD motor distal latency after, left MDMLAR : MD motor distal latency after, right MDMLBL : MD motor distal latency before, left MDMLBR : MD motor distal latency before, right MDPLBR : MD palmar distal latency before, right HANDLR : Left/right hand affected
New Grain Workers' Study (n = 850)	$FEV_1 = \beta_0 + \beta_1\ (base\ height) + \beta_2\ (age)_{ij} + \beta_3\ (weight)_{ij} + \beta_4\ (smoking\ group)_{ij} + \varepsilon_{ij}$
Jaw Growth study (n = 108)	$Y_{ijt} = \alpha_i + \beta_i x_t + \varepsilon_{ijt}$ Where Y_{ijt} - response (distance) for the j^{th} subject in sex group $i = 1, 2$ at time $t = 1,2,3,4$

Table 6.2.: Model concordance correlation coefficient (r_c) ; variance-covariance structure concordance correlation coefficient $r(\phi)$; and pseudo-likelihood ratio test ($\hat{\lambda}$) for four different studies.

		Carpal Tunnel Syndrome Study			
	d.f.	Uncorrelated	Compound symmetric	Unstructured	Autoregressive
r_c		0.4747	0.4747	0.4745	0.4850
$r(\phi)$		0.5884278	0.8525371	0.8069608	0.8338
$\hat{\lambda}$	45	180.57319	93.160274	147.82589	113.1302
		New Grain Workers' study			
	d.f.	Uncorrelated	Compound symmetric	Unstructured	Autoregressive
r_c		0.4308	0.4378	0.6455	0.4340
$r(\phi)$		0.7161062	0.9886752	0.9939072	0.9798073
$\hat{\lambda}$	15	790.2394	27.563722	14.74110	49.25252
		Growth data for the jaws			
	d.f.	Uncorrelated	Compound symmetric	Unstructured	Autoregressive
r_c		0.5943	0.5943	0.5985	0.5955
$r(\phi)$		0.6283748	0.9194793	0.9878345	0.8947285
$\hat{\lambda}$	21	74.775222	14.818215	1.5211502	26.424866

6.2 Previous lung function as a covariate in the Random Effects Models

To determine the behaviour of the goodness of fit statistics for random effects models with previous lung function included as a covariate, random effects models were fitted for the prediction of FEV_1(FVC) with previous FEV_1(FVC) as one of the covariates in addition to all other covariates in the model (5.3). Random effects models were fitted assuming different within subject covariance structures, namely; uncorrelated; compound symmetric; autoregressive; unspecified; unequally spaced AR(1); and unequally spaced AR(1) with observational error. For FEV_1, the between subject covariance matrix was not positive definite for models with uncorrelated, compound symmetric and autoregressive within subject covariance structures, and for unspecified within subject covariance structure restricted maximum likelihood (REML) did not converge. SAS version 6.12 was used to fit the above models, so mixed model equations were not modified, so further calculations needed for the computation of goodness of fit statistics were not done.

In the case of FVC, the between subject covariance matrix was not positive for models assuming uncorrelated, and unequally spaced AR(1) within subject covariance structures. Restricted maximum likelihood did not converge for models with unspecified covariance structures and unequally spaced AR(1) with observational errors. Model concordance correlation coefficients were 0.8880 and 0.8871 respectively, for models assuming compound symmetric; and autoregressive covariance structures. These values were comparatively higher than those obtained for models for FVC when previous FVC was not in the model (Table 5.16), and were closer to the value 1. These results indicate that the random effects models with previous FVC gave

better results than models without previous FVC. The goodness of fit of the covariance structure was not assessed because computation of $r(\hat{w})$ and $\hat{\lambda}$ requires a gold standard covariance structure, which was not possible due to numerical problems.

Based on the model concordance correlation coefficients for FVC models assuming compound symmetric and autoregressive covariance structure, REM with previous lung function seems to give better fit.

PART - II

Outcome variable - Survival time

7. MODELS – LONGITUDINAL MODELS FOR CORRELATED SURVIVAL DATA

7.1 Introduction

Survival data usually arise in clinical and epidemiological prospective studies of humans or laboratory studies of animals. In these studies, time to a clinical response is often the principal response/outcome. The dependence of this univariate measure on covariates come under the topic of survival analysis. In survival analysis, our interest is in a defined point event, called failure, which occur after a certain length of time known as the survival time. Some examples where we can use survival analysis techniques are: i) a study in which we are interested in a first occurrence of a disease after treatment; or ii) a study where we are interested in a death/failure from a lung cancer after an individual was diagnosed with lung cancer (time from diagnoses to death is known as survival time).

Analogous to univariate survival data, we have correlated survival data in longitudinal studies. Correlated survival data can occur in many real-life studies. For such correlated survival data, the variances of estimates of the regression parameters obtained from Cox proportional hazards model are not consistent. So, another method is required to obtain the consistent estimates of variances of regression parameter estimates. In Section 7.2 approaches to the analysis of independent survival data are given and extension of these techniques to handle correlated survival data are given in Section 7.3.

7.2. Survival analysis techniques for cross-sectional independent data

7.2.1. Survival analysis

The survival analysis approach is applicable to certain kinds of data. In survival analysis, a group or groups of subjects is followed for each of whom there is a defined outcome; e.g a disease-free cohort is followed until the occurrence of heart disease. In survival analysis, a few terms e.g. survival time; censoring; survival function; hazard function; and hazard ratio are used very often. These terms and their definitions (Lee, 1992) are described briefly in the next section and details are given by Lee (1992).

7.2.2. Survival Time and Censoring

In survival analysis, the outcome variable is the time until an event occurs, which is usually referred to as the survival time. The survival time can be presented as years, months, weeks, or days from the beginning of follow-up of an individual until an event occurs which is usually referred to as failure. An event can be any specified experience that can occur; e.g. occurrence of death/disease; a specified percent fall in lung function measurement; or relapse from recovery.

Survival data is distinguished from the data arising in other fields of statistics by the presence of censoring, which is a particular form of incomplete data. In survival analysis, censoring is an important concept. Censoring occurs, when we have some information about the survival time, but the exact survival time is not known. Censoring may occur because of the following reasons: 1) an event does not occur for a subject before the study period is over, e.g. some patients may survive to the end of a clinical

trial; 2) a subject withdraws from the study, e.g. a subject who is observed failure-free for a certain time period and then withdraws from a study; or 3) a subject is lost to follow-up during the study.

Mathematically, let T_i be a random variable indicating failure time and C_i be a random variable indicating censored time for an i^{th} individual in a sample of n subjects. Let $X_i = min(T_i, C_i)$, then indicator variable $\delta_i = 1$ if $T_i \leq C_i$ (uncensored, i.e failure occurred) and $\delta_i = 0$ if $T_i > C_i$ (censored). When there are no censored observations, then the data set of survival time is complete. When the subject's exact survival time is incomplete at the end of the follow-up period due to one of the reasons mentioned above, then this kind of data is known as right censored.

7.2.3. Survival function

The survival function estimates the probability that a subject survives longer than some specified time t, i.e.

Survival function $= S(t) =$ P(an individual survives longer than t)

$$= P(T \geq t) \qquad 0 \leq t \leq \infty \qquad \ldots (7.1)$$

Let $F(T)$ be the cumulative distribution function of t. Then

$$S(t) = 1 - P(\text{an individuals fails before time } t)$$

$$= 1 - F(t) \qquad \ldots (7.2)$$

The survival function has the following theoretical properties:

1. Survival functions are monotonically non-increasing.

2. As no event can occur, before the start of the study, therefore at time $t=0$, survivor function takes the value 1, i.e $S(0) = 1$.
3. Theoretically, if the study period as increased to infinity, eventually no one will survive and at $t = \infty$, $S(\infty) = 0$.

The function $S(t)$ is also known as the cumulative survival rate. The graph of $S(t)$ [on y-axis] vs time [on x-axis] is called the survival curve and is used to find the median (50^{th} percentile) survival time.

7.2.4. Probability density function

The survival time T has a probability density function the probability of failure in a small interval per unit time and is defined as the limit of the probability that an individual fails in the short interval t to $t+\Delta t$ per unit width Δt:

$$f(t) = \lim_{\Delta t \to 0} \frac{\text{P\{an individual dying in the interval } (t, t+\Delta t)\}}{\Delta t} \quad \ldots (7.3)$$

The probability density has the following properties:

1. $f(t)$ is a nonnegative function:

$$f(t) \geq 0 \qquad \forall\ t \geq 0$$

$$= 0 \text{ for } t < 0$$

2. the area between the density curve and the *time* axis is equal to 1.

7.2.5. Hazard function

The hazard function of survival time T is denoted by $h(t)$ and gives the conditional failure rate. The hazard function is defined as the failure rate during a small time interval, assuming that the individual has survived to the beginning of the interval, i.e.:

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \le T < t+\Delta t | T \ge t)}{\Delta t} \quad \ldots (7.4)$$

The hazard function can be expressed in terms of the cumulative distribution function $F(t)$ and the probability density function $f(t)$ as:

$$h(t) = f(t) / \{ 1 - F(t)\}$$

$$= f(t)/S(t)$$

The hazard function is also known as instantaneous failure, the conditional mortality rate, the age-specific failure rate or the force of mortality. The hazard function can be graphed as t ranges over various values.

The hazard function may increase, decrease, remain constant or follow any other pattern. The hazard function, $h(t) \ge 0$, has no upper limit. It is not a probability and depends on time units. The cumulative hazard function, $H(t)$ is defined as:

$$H(t) = \int_0^t h(x)dx \quad \ldots (7.5)$$

7.2.6. The Cox's Proportional Hazard Regression Model

The Cox's proportional hazard regression model is widely used in the analysis of censored survival data. It is used for studying the effect of exposure allowing for

confounders and other covariates in cohort studies. It is also used to identify the differences in survival due to treatment and prognostic factors in clinical trials. The Cox's proportional hazard model is briefly defined below:

Let N be the number of individuals in the study, and on i^{th} individual vector $(t_i, \delta_i, \underline{x}_i)$ is observed, where t_i is the time since entry into study; δ_i is the indicator variable of failure, assuming value *1* if the event of interest is observed and *0* if the time is censored; and $\underline{x}_i$ is a covariate vector. The basic Cox's model assumes that the hazard function for failure time T for an i^{th} individual with covariate vector $\underline{x}_i$ (assumed to be constant in time) is

$$h(t, \underline{x}_i) = h_0(t) \exp(\underline{\beta}' \underline{x}_i) \qquad \ldots (7.6)$$

for $i = 1, 2, \ldots, N$. The hazard $h(t, \underline{x}_i)$ given in *(7.6)* depends on both time and covariates. This dependence is through two separate factors $h_0(t)$ *and* $\exp(\underline{\beta}' \underline{x}_i)$. The first factor $h_0(t)$ is a function of time only and assumed to be the same for all subjects; the second factor depends on the individual covariates through vector $\underline{\beta}$ of the regression coefficients.

The hazard ratio for two individuals (i^{th} *and* j^{th}) with covariate vector $\underline{x}_i$ and $\underline{x}_j$. is defined as:

$$\frac{h(t, \underline{x}_i)}{h(t, \underline{x}_j)} = \frac{h_0(t) \exp(\underline{\beta}' \underline{x}_i)}{h_0(t) \exp(\underline{\beta}' \underline{x}_j)} = \exp[(\underline{\beta}' (\underline{x}_i - \underline{x}_j)] \qquad \ldots (7.7)$$

The above model assumes that the failure rates of any two individuals are proportional, given that the ratio in *(7.7)* does not depend on time. The hazard function $h_0(t)$ is known as the baseline hazard, because this can be regarded as a hazard function of an individual with zero values for all the covariates.

One of the goals of survival analysis is to assess the relationship of the explanatory variables to survival time. Survival analysis techniques have been used to analyse the bronchial hyperresponsiveness data. The next section shows how one can consider bronchial hyperresponsiveness measurements as survival data.

7.2.7. Hazard Ratio

The hazard ratio is an instantaneous relative risk of an event per unit time for an individual with the risk factor present compared with an individual with the risk factor absent, given that both individuals have survived to time t. For a dichotomous independent variable, the hazard ratio is given by:

$$h(t/x_i=1)/h(t/x_i=0) = exp(\beta_i) \quad ... (7.8)$$

For a continuous independent variable, the hazard ratio is given by:

$$h(t/x_i+\Delta)/h(t/x_i) = exp(\beta_i\Delta) \quad ... (7.9)$$

7.3 Bronchial responsiveness measurements as censored survival data

In measurements of bronchial hyperresponsiveness data, subjects who do not reach a 20% decrease in FEV_1 (for definition of FEV_1, see Appendix B) after the maximum dose of histamine, PC_{20} (for definition of PC_{20}, see Appendix B) considered to be the maximum dose/concentration for statistical calculations. As in these subjects PC_{20} known to exceed the maximum dose, the traditional analysis underestimates the effect of stimulus. We propose an alternative method to analyse these kind of data using survival analysis techniques.

According to the definitions of survival time and censoring time given above,

measurements of bronchial hyperresponsivness can be considered as censored survival data. In measurements of bronchial hyperresponsiveness, failure time is defined as the dose at which a 20% decrase in FEV_1 occurred. The histamine dose is considered equivalent to "time", and if the $PC_{20} \leq 8$ mg, then it is defined as a "failure". The exact value of PC_{20} was calculated by linear interpolation between the last two points. The concentration of histamine was plotted against the percent decrease in FEV_1. The following formula (Cockroft, 1977; and Hargreave et al., 1981) was used for the linear interpolation of the PC_{20} from the log dose-response curve:

$$PC_{20} = antilog\left[logC1 + \frac{(logC2 - logC1)(20 - R1)}{(R2 - R1)}\right] \qquad \ldots (7.10)$$

where:

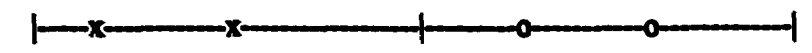
C1= second last concentration (<20 % FEV_1 decrease)

C2 = last concentration (> 20 % FEV_1 decrease)

R1 = % decrease in FEV_1 after *C1*

R2 = % decrease in FEV_1 after *C2*

Censoring occurs when the histamine dose required for a 20% decrase in FEV_1 exceeds 8 mg and the exact dose is not known. The concepts of failure and censoring are shown using a following diagram:



0 mg 8 mg

In the above diagram x denotes a failure and o denotes censoring, i.e. subjects were administered doses only up to 8 mg and were censored at 8 mg.

Survival analysis techniques are unique in the ability to incorporate censoring in statistical analysis. The Cox proportional hazard model was used to analyze bronchial hyperresponsiveness data available on new grain workers and to examine the association between explanatory variables and hyperresponsiveness after allowing for censoring.

7.4. Extension of proportional hazard models to correlated survival data

In the new grain workers' study, explained in Section 3.3, Chapter 3, we have at most four observations per individual. These data for each individual are correlated. We extended the survival analysis approach (Cox regression model) to correlated survival data. An important assumption of the Cox model is that the failure times are mutually independent. This assumption is violated, if there are some correlated events. In the new grain workers' study, there are correlated events because of the repeated observations on each subject, so the assumption of mutually independent failure times is violated. Due to this correlation, the variances of the estimates of regression coefficients obtained from the usual analysis are not valid in these data, so we used other techniques e.g. jackknife and bootstrap discussed in Section 7.5, to obtain the valid variance estimates.

In our data analysis, the i^{th} subject will be treated as the i^{th} cluster (each cluster will be treated as a stratum), and

1. Let T_{ik} be the failure time at the k^{th} observation (we have four observations for each subject) for i^{th} subject (or stratum)
2. Let $(C_{i1}, C_{i2}, C_{i3}, C_{i4})'$ be the censoring vectors.
3. Let $X_{ik} = min(T_{ik}, C_{ik})$ and $\delta_{ik} = 1$ if $X_{ik} = T_{ik}$ and $\delta_{ik} = 0$ otherwise.

4. Let $Z_{ik}(t)$ be a $p \times 1$ vector of covariates for T_{ik} at time $t>0$.

We have the made the following assumptions for fitting a Cox proportional hazard model for correlated censored survival data:

1. The failure times within cluster are assumed to be independent.
2. The number of observations within a cluster/stratum (subject) are relatively small compared to the number of clusters/strata (subjects).
3. The censoring vectors are independent of the failure time vectors.

The hazard function for T_{ik} at time $t > 0$ has the usual proportional hazards form (Lipsitz and Parzen, 1996):

$$h(t \mid Z_{ik}(t)) = h_0(t) \exp(\underline{\beta}' Z_{ik}(t)) \quad t>0 \qquad ...(7.11)$$

where $h_0(t)$ is an arbitrary hazard function and $\underline{\beta}$ denotes the vector of the true regression coefficient. We are interested in an estimator $\hat{\underline{\beta}}$ of $\underline{\beta}$ which maximizes the logarithm of the "partial likelihood" function (Wei et al., 1989; and Lee et al. 1992; Lipsitz and Parzen, 1996). Although, the individual observations in each subject/stratum are correlated, $\hat{\underline{\beta}}$ is a consistent estimator for $\underline{\beta}$, but the variance estimates of the regression parameters obtained from the Cox model are not valid estimates. So, other methods e.g given by Wei et al (1989), generally known as WLW, jackknife (Lipsitz et al, 1994; and Lipsitz and Parzen, 1996) or bootstrap were used to obtain the consistent estimators of variances for the regression parameters.

7.5. Jackknife and bootstrap techniques

Jackknife and bootstrap techniques are closely related inferential techniques. Both techniques assess the variability of a statistic by examining the variation within the sample data, rather than through the use of parametric assumptions. The jackknife technique consists of forming new samples by dropping, in turn, one of the observations of the original sample. For each of the samples thus generated, the estimator under study is calculated, and then the variation in the estimator is assessed by calculating the standard error. The jackknife algorithm for estimating standard error for independent observations can be found in Efron and Tibshirani (1997).

In longitudinal studies, where observations are correlated, a cluster(subject) may consist of a single observation or more depending on a number of repeated measurements taken on a subject. When using the jackknife technique for correlated data, a cluster is dropped out one at a time and Cox's proportional hazards model is fitted to estimate the relative risk parameters. The jackknife estimator of variance of $\hat{\underline{\beta}}$ was computed by using the formula (Lipsitz et al., 1994; and Lipsitz and Parzen, 1996):

$$\hat{Var}(\hat{\underline{\beta}}) = \left(\frac{N-p}{N}\right)\sum_{i=1}^{N}(\hat{\underline{\beta}}_{-i} - \hat{\underline{\beta}})(\hat{\underline{\beta}}_{-i} - \hat{\underline{\beta}})' \qquad ...(7.12)$$

where $\hat{\underline{\beta}}_{-i}$ is the estimate of $\underline{\beta}$ obtained by deleting the n_i^{th} observation in cluster i.

As suggested by Lipstiz et al (1994), comparison of performance of bootstrap estimates of variance to jackknife needs further investigation. We obtained the bootstrap estimate of variance for βs obtained by using Cox's partial likelihood method. The bootstrap technique consists of drawing with replacement many new samples of the

same size or less as the observed sample, by drawing a random selection of the original observations, i.e. drawing some of the observations with replacement several times. The estimator under study is calculated for every one of the samples thus generated, and will show a probability distribution of its own. From this distribution, the standard error of the estimator under study is computed.

In the bootstrap technique for correlated censored survival data, 1000 random samples of size N from the original N clusters with replacement were drawn and a Cox's proportional hazards model was fitted to estimate the regression parameter, β. This technique was suggested by Kunsch (1989) who used the bootstrap method for time series by resampling blocks, where a block is defined as a group of consecutive observations in the time series. Bootstrap estimates of regression parameters and their variances are obtained by taking the mean and variance of 1000 regression estimates obtained by repeating the resampling 1000 times.

The bootstrap estimate of the standard error by the sample standard deviation of the B replications is:

$$\hat{se}_B = \left\{ \sum_{b=1}^{B} \left[\hat{\beta}(b) - \underline{\hat{\beta}}(.) \right]^2 / (B-1) \right\}^{1/2}$$

where $\hat{\beta} = \sum_{b=1}^{B} \hat{\beta}(b)/B$ and $\hat{\beta}(b)$ is the estimate of the parameter of interest obtained from the b^{th} bootstrap sample (b = 1, 2, ... , B).

We fitted Cox's regression model for correlated censored survival data, which produced consistent estimators for the regression coefficients. A SAS macro was developed to compute the jackknife and bootstrap estimators of the standard errors of

the parameters. We computed a robust asymptotic estimator proposed by Wei et al for variance of regression parameters by using a SAS program PHRWLW (available on SAS SAMPLE LIBRARY). We also compared this robust asymptotic estimator with jackknife and bootstrap estimators.

8. AN APPLICATION OF THE PROPORTIONAL HAZARD MODEL TECHNIQUES TO LONGITUDINAL DATA TO DETERMINE THE PREDICTORS OF THE FIRST EPISODE OF WHEEZING[@]

8.1 Introduction

In this chapter, the proportional hazard model technique is used to analyze the longitudinal data in order to determine the predictors of the first episode of wheezing among grain elevator workers. We report the results of a longitudinal study of Canadian grain elevator workers, in which we determined the predictors of a first episode of wheezing in apparently healthy and asymptomatic grain workers at baseline, and for whom we have at least two sets of observations. We report estimates of the magnitude of risk of developing a first episode of wheezing for each predictor after adjusting for other factors.

8.2 Study subjects and statistical methods

A respiratory health surveillance program for grain elevator workers in Canada commenced in 1978 (details of this program are given in Section 3.2, Chapter 3). Grain elevator workers with abnormal chest X-rays and/or with presence of any respiratory symptoms (wheeze, dyspnea, cough or sputum) and/or asthma at baseline [Cycle II], were not included in the analysis reported in this chapter. The information on wheezing

[@] *This chapter is based on the paper "Predictors of onset of wheezing in grain elevator workers" published in the Canadian Respiratory Journal, 1998, 5 (3), 200 - 205.*

was based on responses to two questions: "Does your chest ever sound wheezing or whistling?" and "Do you get this most days and nights?" The presence of dyspnea was determined from the question: "Are you troubled by shortness of breath when hurrying on level ground or walking up a slight hill?" The presence of chronic cough and phlegm was determined if the symptom was present in the morning or during the day or night for more than three months a year for two years. The development of a first episode of wheezing was the main interest of the present investigation, and only asymptomatic subjects at baseline were included in the analysis. Data from one province did not conform to the standards established by Labour Canada (Pahwa et al., 1994), and therefore these data were not included in the analysis. Data from Cycle I were not included in this analysis because the follow-up between Cycle I and Cycle V was incomplete.

At the baseline (Cycle II), 5493 male grain workers from 27 grain elevator companies participated in the study. Complete information was available on 4671 subjects. Of these subjects, 1919 participated only in Cycle II and were excluded from this analysis. Significant differences were observed in age, height, FVC, FEV_1, ratio of FEV_1/FVC and smoking between subjects who were studied again in Cycle III, Cycle IV or Cycle V (n=2752) and those who were excluded (n=1919). Among the former group (n=2752), 1848 subjects were symptom free (asymptomatic) and 904 subjects (symptomatic) reported one or more respiratory symptoms (n=903), had abnormal X-rays (n=82) or had physician-diagnosed asthma (n=6). Table 8.1 shows the comparison of the baseline characteristics between asymptomatic (n=1848) and symptomatic

(n=904) subjects. Symptomatic subjects were significantly older, had longer duration of employment, were currently smoking more and had lower mean values of pulmonary function measurements in comparison to the asymptomatic subjects. Only the asymptomatic subjects were considered for the analysis to determine the predictors of a first episode of wheezing.

Smoking behaviour was determined at baseline and at the end point (Senthilselvan et al, 1996). Subjects who were nonsmokers at baseline and at the end point were allocated to the lifetime "nonsmoker" category. The "ex-smoker" category was comprised of (i) subjects who reported smoking at the baseline and denied smoking at the end point; (ii) subjects who were exsmokers at the baseline and at the end point; and (iii) subjects who were exsmokers at the end point. Subjects who reported smoking at the end point were allocated to the "current smoker" category.

Statistical Methods

The annual rate of decline in lung function was calculated for each subject by dividing the difference between the baseline and endpoint lung function measurements by the time period between the baseline and endpoint. Baseline characteristics of asymptomatic subjects who developed wheezing during the period 1981-1993 were compared with those subjects who did not develop wheezing during this period. Two sample t-tests were used to compare the continuous variables age, FEV_1, FVC, FEV_1/FVC ratio, and duration of follow-up. The Chi-square test was used for comparisons of the categorical variable smoking behaviour.

Table 8.1 : Baseline characteristics of asymptomatic and symptomatic grain workers who continued to work at the beginning of Cycle II.

	Asymptomatic at Cycle II (n=1848)	Symptomatic at Cycle II (n=904)
	Mean (SD)	Mean (SD)
Age, yr	34.0 (11.4)	38.3 (12.2)*
Years in industry	9.9 (8.7)	13.7 (9.8)*
FEV_1, *l*	4.2 (0.7)	3.9 (0.8)*
FVC, *l*	5.3 (0.8)	5.1 (0.9)*
FEV_1/FVC	80.2 (6.6)	76.6 (8.1)*
	n (%)	n (%)
Smoking status		
Never smokers	635 (34.4)	151 (16.7)
Exsmokers	526 (28.5)	204 (22.6)
Current smokers	687 (37.2)	549 (60.7)[]

* $p<0.0001$

[] $p<0.001$

Survival analysis techniques were used to determine the risk factors for a first episode of wheezing. In using survival analysis, we defined the time of origin, end point, survival time, censoring and censoring time. The time of origin for this study was Cycle II and the end point was either Cycle III, Cycle IV or Cycle V depending on when the subject last participated in the surveillance program. Survival time was the period from Cycle II to the Cycle at which the subject reported an episode of wheezing. Censoring occurs in our study when a subject did not report an episode of wheezing at any cycle. Censoring time was defined as the period from Cycle II to the last cycle in which the subject participated without a report of wheezing.

The variables considered in the analysis were age, exposure years in the grain industry, height, smoking behaviour and FEV_1/FVC ratio at baseline. Exposure years in the grain industry were divided into three categories: (i) < 10 years; (ii) ≥ 10 and < 20 years and (iii) ≥ 20 years. Two dummy variables for exposure years were used in the analysis.

Cox's proportional hazards model (Lee, 1992) is of the form:

$$log[h(t|x)/h_0(t)] = exp(\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7)$$

where $h(t|x)/h_0(t)$ is the hazard ratio and seven variables in the model represent age at baseline (x_1), dummy variable for exposure years ≥ 10 and < 20 (x_2), a dummy variable for exposure ≥ 20 years (x_3), height at baseline (x_4), a dummy variable for ex-smokers (x_5), a dummy variable for current smokers (x_6) and the ratio FEV_1/FVC at baseline (x_7). One important assumption of the hazards model is that different individuals have hazard functions that are proportional to one another. The validity of the assumptions

made in Cox's model was tested by fitting the interaction between time and the significant predictor variables (age and FEV_1/FVC ratio) one at a time (Lee, 1992). For example, to test the proportionality assumption for age, we fitted a Cox's model with age and the product of age and survival time as predictor variables.

Survival analysis techniques were used in a similar manner as above to examine the predictors for a first episode of dyspnea, cough, or sputum.

8.3 RESULTS

Table 8.2 shows the proportions of symptoms (wheeze, dyspnea, cough and sputum) in Cycle III, Cycle IV and Cycle V among 904 workers who were symptomatic at Cycle II. At Cycle II, 41.5% reported wheeze, followed by 34.6% at Cycle III, 13.9% at Cycle IV and 11.0% at Cycle V. Fewer subjects reported dyspnea at Cycle II (35.2%) and Cycle III (22.9%). There was a decreasing trend for all symptoms over the study period suggesting a "healthy worker" effect. Table 8.3 shows the distribution of the first episode of wheezing and censoring during the study period from Cycle II (1981-84) to the end of the study.

Table 8.4 shows demographic, smoking, and lung function test values in the asymptomatic grain workers at baseline, who subsequently reported an episode of wheezing during the study period, compared with those who remained wheeze free. The proportion of current smokers was significantly higher among those who reported wheezing (48.8%) than among those who did not report wheezing (29.1%). The baseline and endpoint FEV_1/FVC ratios were significantly higher for those who did not report wheezing compared to those who reported an episode of wheezing. Crude

annual declines in FEV_1, FVC and the FEV_1/FVC ratio were significantly higher for those who reported wheezing compared to those who did not report wheezing. Similar trends were observed for annual declines in FVC and of the FEV_1/FVC ratio.

Table 8.5 shows the results of fitting Cox's proportional hazards model to identify predictors for the development of wheezing, dyspnea, cough and sputum in grain workers. As age and years in the grain industry were highly correlated (Pearson's correlation coefficient=0.776, $p<0.001$), models were fitted with and without including age. When age was included in the model, the relative risk for years in the industry was reduced, so age was kept in the model. The baseline FEV_1/FVC ratio, and smoking behaviour during the study period were significant predictors of a first episode of wheezing after adjusting for height. Those who had a higher FEV_1/FVC ratio at baseline experienced a protective effect [relative risk (RR): 0.02; 95% confidence interval (C.I): (0.003, 0.20)] against developing wheezing in the multivariate model. Current smokers were at increased risk [RR: 2.33, 95% C.I: (1.63 - 3.33)] of developing wheeze after controlling for age, height, the FEV_1/FVC ratio and exsmoking status.

Significant predictors for first episode of dyspnea were age and current smoking. Significant predictors of a first episode of cough or first episode of sputum were current smoking and baseline FEV_1/FVC ratio. A similar analysis was conducted to examine the predictors for development of both wheezing and dyspnea in grain workers. Years in the industry, current smoking and the FEV_1/FVC ratio were significant predictors of first episodes of wheezing and dyspnea in the absence of age in the model.

A dose-response relationship was also observed between years in the industry and the first episode of wheezing and dyspnea. When age was included, years in the

Table 8.2 : Proportions of respiratory symptoms among grain workers who were symptomatic at Cycle II (n=904).

	Cycle II (n=904)		Cycle III (n=691)		Cycle IV (n=366)		Cycle V (n=337)	
	n	%	n	%	n	%	n	%
Wheeze	375	(41.5%)	239	(34.6%)	51	(13.9%)	37	(11.0%)
Dyspnea	318	(35.2%)	158	(22.9%)	60	(16.4%)	44	(13.1%)
Cough	441	(48.7%)	244	(35.3%)	105	(28.7%)	83	(24.6%)
Sputum	439	(48.5%)	256	(37.1%)	95	(26.0%)	67	(19.9%)

Table 8.3 : Distribution of first episode of wheezing and censoring during the follow-up.

	Baseline (Cycle II)	Cycle III	Cycle IV	Cycle V	
Follow-up	Symptom Free	Wheeze	Wheeze	Wheeze	Censoring at end Point
Cycle II to Cycle V	808	46	17	32	713
Cycle II to Cycle IV	358	25	14	-	319
Cycle II to Cycle III	682	69	-	-	613
Total:	1848	140	31	32	1645

Table 8.4 : Characteristics of grain workers who reported wheezing compared to those who did not report wheezing during the study period.

	Wheezing reported	No wheezing reported
	Mean (SD) (n=203)	Mean (SD) (n=1645)
Age, yr	35.5 (11.4)	33.8 (11.4)
Height, cm	175.6 (6.8)	174.9 (6.5)
Duration of follow-up, yr	4.4 (2.1)	6.0 (2.6)[¤]
FEV_1		
Baseline (*l*)	4.15 (0.8)	4.23 (0.7)
End point (*l*)	3.85 (0.9)	4.04 (0.8)[¤]
Annual rate of decline (*ml*)	85.3 (143.3)	39.3 (97.4)[¤]
FVC		
Baseline (*l*)	5.32 (0.9)	5.26 (0.8)
End point (*l*)	5.00 (1.0)	5.08 (0.9)
Annual rate of decline (*ml*)	88.2 (155.1)	40.0 (114.7)[¤]
FEV_1/FVC		
Baseline (%)	78.0 (7.1)	80.5 (6.5)[¤]
End point (%)	76.6 (8.5)	79.7 (7.0)[¤]
Annual rate of decline (%)	0.4 (1.6)	0.1 (1.2)[¤]
	No (%)	No (%)
Smoking behaviour[@]		
Nonsmokers	45 (22.2)	539 (32.8)
Exsmokers	59 (29.1)	627 (38.1)
Current Smokers	99 (48.8)	479 (29.1)[*]

[*] $p<0.0001$; [¤]: $p < 0.01$; [¤]: $p < 0.001$; [¤]: $p < 0.05$ [@] Smoking behaviour was calculated from all cycles (Please see methods section)

industry was not significant. Similar analyses were conducted for first episode of cough and sputum together. Years in the grain industry was not a significant predictor for a first episode of cough, sputum, or cough and sputum. Smoking and the FEV_1/FVC ratio were significant predictors for cough and sputum occurring together.

We used "years in the grain industry" as a surrogate for personal dust exposure. However, personal dust samples were available for 17 of the grain elevator companies that participated in the study. The dust samples were collected by the Labour Canada Regional Inspectional staff in response to specific complaints and at the discretion of the inspectors. A total of 340 grain dust samples from 14 terminal elevators and 190 samples from 3 primary elevators were available from 1980 to 1984. No identifiers were available, thereby maintaining confidentiality. Therefore, we were not able to match dust level measurements data with symptoms and lung function test values. However, analysis of the 530 grain dust samples showed that workers in certain job classifications were exposed to higher dust levels than were others. Based on this analysis three job categories were created: (i) mean dust levels > 10 mg/m^3; (highest exposed group) (ii) mean dust levels ≤ 10 mg/m^3 and > 5 mg/m^3; (moderately exposed group), and (3) mean dust levels ≤ 5 mg/m^3; (lowest exposed group). We allocated the 1849 workers from the 27 companies to the three job categories based on the information that we could obtain about these companies. This categorization was used as a surrogate index of dust exposure. When we included two dummy variables as indicators of the three job categories in Cox's model with other predictors, the dummy variables were not statistically significant.

Figure 8.1 shows wheeze-free survival probability curves stratified by smoking behaviour and by the FEV_1/FVC ratio. Nonsmokers with $FEV_1/FVC \geq 70\%$ had the lowest risk of developing wheeze during the study period and current smokers with $FEV_1/FVC < 70\%$ had the highest risk of developing wheeze. The validity of the proportionality assumption of the predictor variables age and ratio of FEV_1/FVC were tested and no major violations were found.

8.4 DISCUSSION

In this chapter, longitudinal data was treated as a survival data and analyzed by using a one of the popular mathematical models, Cox proportional hazard model. The outcome of interest was the first episode of wheezing. The objectives of the statistical analysis in this chapter were to observe determinants of a first episode of wheezing; and to determine the magnitude of the risk of developing a first episode of wheezing for particular predictors adjusting for other factors by using the Cox's proportional hazard model. Several conclusions can be drawn from this analysis of longitudinal data of grain workers who participated in more than one surveillance cycles over a 12 year period. The group of nonsmokers with FEV_1/FVC ratio greater than or equal to 70% had the best survival function for not developing wheeze as compared to nonsmokers with FEV_1/FVC less than 70% and smokers with FEV_1/FVC greater or less than 70%. The risk of risk of a first episode wheezing also increased with age.

The incidence of first wheeze was highest in Cycle III (the first cycle after the baseline) and was followed by a decreasing trend. A similar trend was observed for the incidence of wheezing and dyspnea. This could be the result of a number of factors.

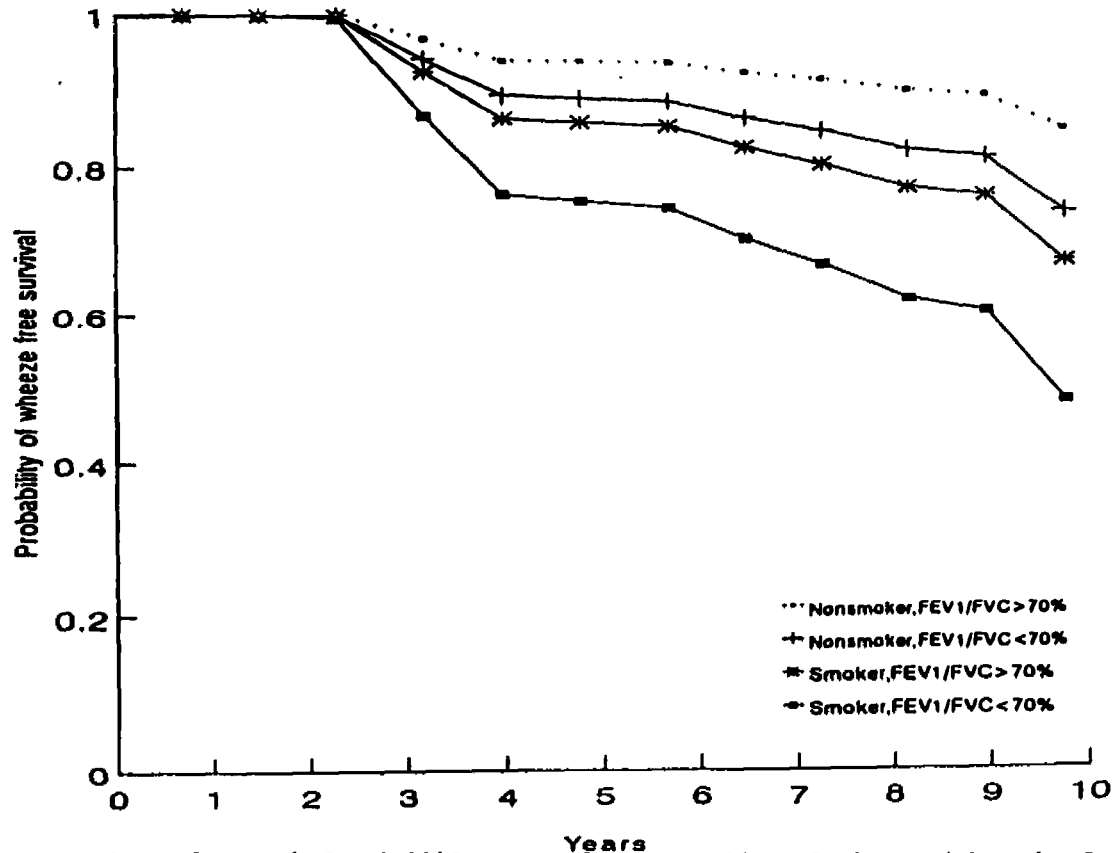


Figure 8.1. Wheeze-free survival probabilities categorized by smoking behaviour and the ratio of forced expiratory volume in 1 second to forced vital capacity(FEV_1/FVC).

Table 8.5 : Proportional hazard regression analysis to identify predictors for development of any symptom in grain workers.

	Symptom							
	Wheeze		Dyspnea		Cough		Sputum	
Variable	Odds ratio	95% C.I.	Odds ratio	95% C.I.	Odds ratio	95% C.I.	Odds ratio	95% C.I.
Age	1.01	1.00 - 1.03	1.06	1.04 - 1.09	1.01	1.00 - 1.03	1.02	1.00 - 1.04
Yrs in industry:								
≥10 and <20 yrs	1.06	0.73 - 1.52	1.06	0.68 - 1.66	1.00	0.68 - 1.45	0.86	0.58 -1.28
≥20 yrs	1.06	0.62 - 1.83	1.23	0.70 - 2.17	0.82	0.44 - 1.52	0.88	0.48 - 1.61
Height	1.02	0.99 - 1.04	1.02	1.00 - 1.05	1.00	0.98 - 1.02	1.01	0.99 - 1.03
Exsmoker	0.93	0.62 - 1.38	1.11	0.68 - 1.81	1.14	0.70 - 1.86	1.16	0.73 - 1.03
Current smoker	2.33	1.63 - 3.33	2.15	1.32 - 3.48	4.68	3.08 - 7.11	3.63	2.42 - 5.46
FEV_1/FVC	0.02	0.003 - 0.20	0.17	0.01 - 2.56	0.06	0.006 - 0.62	0.07	0.007 - 0.78

There may have been a healthy worker effect (Broder et al., 1985) in these populations, and the population that was followed-up after Cycle III might be healthier than those who were lost to follow-up. Other factors may include improved working conditions and job seniority, which could lead to less dusty working conditions for more senior workers.

The association between first onset of respiratory symptoms and occupational exposure has been investigated in other industries (Kongerud and Samuelsen, 1991). Kongerud and Samuelson (1991), in a prospective study of respiratory health in aluminium potroom workers found that 8.1% of workers reported dyspnea and wheezing during follow-up. Our results showed that 11.0% of asymptomatic grain workers developed wheezing during an observation period of nine years. During the study period, 11.5% of grain workers reported wheezing or dyspnea and 3.6% reported a combination of wheezing and dyspnea (data not shown). Konegerud and Samuelsen (1991) used a proportional hazards analysis to determine the predictors of the development of dyspnea and wheezing. They found that smoking and total fluoride exposure were the most important predictors. They reported that the risk of developing dyspnea and wheezing among smokers was two to three times higher than that for nonsmokers. Our data show that the risk of developing wheezing among current smokers was 2.3 times that of nonsmokers.

In a prospective study of middle-aged and older men who initially denied any history of wheezing and asthma, current smoking was the strongest independent predictor of new onset of wheezing (Sparrow et al, 1993). We also found that current smoking was a risk factor for a first episode of wheezing. McDuffie et al (1991)

reported on the respiratory health status of 3098 Canadian grain workers studied longitudinally at two different time points; 1981-84 (Cycle II) and 1984-87 (Cycle III). The frequency of chronic sputum production and chronic wheeze changed significantly from Cycle II to Cycle III. Obstructive lung dysfunction was more prevalent and increased from Cycle II to Cycle III. In the present analysis, we found that grain workers with lower lung test values were at increased risk of developing wheeze.

Our analysis was aimed at evaluating predictors for the development of wheeze among initially symptom free grain workers studied longitudinally. This study provides evidence that among Canadian grain elevator workers, independent predictors of future development of wheezing are current smoking and the baseline FEV_1/FVC ratio; independent predictors of future development of dyspnea are age and current smoking; independent predictors of future development of cough are current smoking and baseline; and independent predictors of sputum are current smoking and baseline FEV_1/FVC ratio.

Cox's proportional hazard model, a multivariate technique proved to be useful in investigating the relationship between survival time (time to first episode of wheezing) and possible prognostic variables.

9. RESULTS AND DISCUSSION – PROPORTIONAL HAZARDS MODELS FOR CORRELATED SURVIVAL DATA

9.1 Introduction

In provocation tests, the end point is the provocation dose (PD_{20}) or the provocation concentration (PC_{20}) required for a 20% decrease in the forced expired volume in the first second (FEV_1). In the longitudinal data set of new grain workers considered in this chapter, there were at most four observations per subject. Survival analysis techniques were initially used to analyse baseline (cross-sectional) data to determine the predictors for bronchial hyperresponsiveness (BHR). Finally, Cox's proportional hazards models for correlated survival data were used to analyze grain workers' longitudinal data.

To our knowledge, survival analysis techniques have not been previously used for correlated bronchial hyperresponsiveness data. Using these techniques, we have attempted to evaluate the significance of factors that are associated with airway hyperresponsiveness. Cox's partial likelihood approach was used to estimate the risk parameters and jackknife, bootstrap, and WLW methods to determine the robust asymptotic variances of the parameter estimates. Jackknife, bootstrap, and WLW robust asymptotic variance estimators were compared with those obtained from the usual partial likelihood method.

For comparison with the results of the proportional hazards model, repeated logistic regression analysis was also conducted to examine the predictors of BHR.

9.1 Study subjects

The data used in this chapter was described in Section 3.2, Chapter 3. We studied 217 men who had just commenced work in the grain industry and 118 male control subjects (baseline). Bronchial responsiveness measurements were not available for two grain workers at the baseline. One control refused histamine challenge tests at the 1st and 2nd recall. The data from these three subjects (four observations) were excluded from the statistical analysis. The number of grain workers and control subjects who had valid histamine challenge test measurements at different time points are given in Table 9.1.

Table 9.1 : Number of grain workers and control subjects examined at different time points.

Study	Grain Workers	Controls	Total
Baseline	217	118	335
1st recall	117	100	217
2nd recall	108	97	205
3rd recall	53	40	93
Total number of observations	495	355	850

9.3. Statistical Analysis

Descriptive results for continuous variables, e.g age, packyears, exposure to grain dust (in weeks), FEV_1 and FVC were expressed as mean ± standard errors. Categorical variables (e.g. positive skin test ; smoking status; wheezing; and occurrence of PC_{20}) were described with frequencies and percentages. Chi square tests were used to determine the association between PC_{20} and other factors. If the expected value was less than 5, Fisher's exact test was used instead of the Chi square test (Colton, 1974).

Cox's proportional hazards model was fitted to baseline data to determine the predictors for bronchial hyperresponsiveness:

$$log[h(t|x)/h_0(t)] = exp(\beta_1 * group + \beta_2 * skin\ test + \beta_3 * FEV_1 + \beta_4 * smoking\ status + \beta_5 * wheezing + \beta_6 * age + \beta_7 * height) \quad \ldots (9.1)$$

We compared the results of utilizing the Cox regression model with those obtained using logistic model. In logistic regression, we considered a dichotomous outcome, i.e. whether PC_{20} has occurred or not occurred at a histamine dose of 8 mg/ml. A logistic model can be represented as:

$$ln[p(x)/(1-p(x)] = \beta_0 + \beta_1 * group + \beta_2 * skin\ test + \beta_3 * FEV_1 + \beta_4 * smoking\ status + \beta_5 * wheezing + \beta_6 * age + \beta_7 * height \quad \ldots (9.2)$$

where $p(x)$ is probability of an event occurring in an interval during which measurements were obtained. An important difference between *(9.1)* and *(9.2)* is that the logistic model assumes that no persons were lost during the follow-up period, while the Cox model accounts for such individuals.

The Cox regression model considered for the correlated survival data is:

$$\lambda_0(t) \exp [\beta_1 * group + \beta_2 * (skin\ test)_{ij} + \beta_3 * (FEV_1)_{ij} + \beta_4 * (smoking\ status)_{ij}$$

$$+ \beta_5 * (wheezing)_{ij} + \beta_6 * (age)_{ij} + \beta_7 * (height)_{ij}] \quad \ldots (9.3)$$

Details of an extension of the Cox proportional hazard model to correlated survival data were given in Section 7.4, Chapter 7. A comparison was made between the standard errors obtained from the partial likelihood with those obtained from jackknife and bootstrap methods. A SAS macro was developed to compute the jackknife and bootstrap estimators of the standard errors of the parameters (SAS Guide to Macro Processing). We also computed WLW robust asymptotic estimator (Wei et al, 1989) and compared it with jackknife and bootstrap estimators.

The results of correlated survival data analysis were compared with those from logistic regression for repeated binary data by using the Generalized Estimating Equation approach introduced by Liang and Zeger (1986). The logistic marginal model for longitudinal data is of the form:

$$logit(\mu_{ij}) = \beta_0 + \beta_1 * group + \beta_2 * (skin\ test)_{ij} + \beta_3 * (FEV_1)_{ij} + \beta_4 * (smoking\ status)_{ij} +$$

$$\beta_5 * (wheezing)_{ij} + \beta_6 * (age)_{ij} + \beta_7 * (height)_{ij} + \varepsilon_{ij} \quad \ldots (9.4)$$

and we also assumed that $Var(Y_{ij}) = \mu_{ij}(1 - \mu_{ij})$ and $Corr(Y_{ij}, Y_{ik}) = \alpha$, where α is the within subject correlation between the j^{th} and k^{th} observation for the i^{th} subject. Details about repeated logistic regression can be found in Section 4.5.2, Chapter 4.

9.4 Results

The mean (± standard deviation) of age; lung function values; and presence of positive skin test and bronchial hyperresponsiveness among stay-in and drop-out subjects are shown in Table 9.2. Nearly 61.8% (n=134) of the grain workers remained

in the study for future recalls after the baseline evaluation. Of these 134 grain workers, 117 grain workers were studied at the 1st recall, and the remaining 17 grain workers returned for participation at the 2nd recall. Of 117 grain workers studied at the 1st recall, 77.8% (n=91) stayed-in for the 2nd recall and of these, 58.2% stayed-in for the 3rd recall. Among controls 85.6% stayed-in for the 1st recall and of those who were studied at the 1st recall, 92% were studied at the 2nd recall. Of these, 40.7% were studied at the 3rd recall.

Table 9.3 shows the descriptive statistics of demographic characteristics and lung function values of cases and controls at baseline and at each of the three recalls. There was a substantial drop out of subjects from the 2nd to the 3rd recall. By the 2nd recall, grain workers had significantly lower FVC compared to controls. Grain workers had significantly lower FEV_1 at baseline and at the second and third recalls in comparison to the controls.

Table 9.4 shows the prevalence of wheezing, positive skin test, bronchial hyperresponsiveness and smoking status among grain workers and control subjects at the four time points. The percentage of wheezing increased in grain workers and controls from baseline to the 3rd recall. There were significantly more smokers among grain workers and the percentage of exsmokers increased from baseline to the 2nd recall among grain workers.

Cox's regression model (9.1) and the logistic model (9.2) were fitted for the baseline data (217 grain workers and 118 controls). Results of survival and logistic analyses for baseline data are shown in Table 9.5 and Table 9.6 respectively. Both these analyses show that FEV_1 was the significant predictor of bronchial

hyperresponsiveness in the multivariate model, indicating that subjects with lower values of FEV_1 were at higher risk of experiencing a 20% fall in FEV_1.

In Tables 9.7 and 9.8, we present the results of using Cox's proportional hazard model for correlated observations. The predictors considered in the cross-sectional analysis were re-examined in the longitudinal analysis for correlated censored data. To take into account the correlation among repeated observations on each subject and to obtain consistent estimators of standard errors, we used jackknife, bootstrap and WLW methods to estimate the variances of parameter estimates. The partial likelihood, WLW, jackknife, and bootstrap estimators of standard errors are shown in Table 9.8. The risk ratios and 95% confidence intervals (CI) computed by using these four methods are shown in Table 9.8. The results obtained from applying the partial likelihood method (which naively assumes that observations within subject are independent) were different from those using robust methods. Based on the partial likelihood method, the significant predictors were positive skin test, low FEV_1, wheezing and height. Based on the three robust methods, the significant predictors of bronchial hyperresponsiveness were: positive skin test, low FEV_1, smoking status and wheezing. Smoking status was not significant based on the partial likelihood method.

The results of the repeated logistic regression analyses (based on GEE approach) with unstructured covariance structure for within subject dependencies are given in Table 9.9. The significant predictors based on this method were: low FEV_1 and height. These results were different from those obtained by survival correlated data analyses based on the partial likelihood method, and the other three robust methods. Assuming the standard errors obtained from utilizing the robust methods are correct (Lipsitz and

Table 9.2 : The distribution of Age, lung function tests, allergy and bronchial hyperresponsiveness in stay-in and drop-outs

	Baseline to 1st recall				1st recall to 2nd recall				2nd recall to 3rd recall			
	Grain workers		Controls		Grain Workers		Controls		Grain workers		Controls	
	Stay-in (n=134) Mean±SD	Drop-out (n=83) Mean±SD	Stay-in (n=105) Mean±SD	Drop-out (n=13) Mean±SD	Stay-in (n=91) Mean±SD	Drop-out (n=26) Mean±SD	Stay-in (n=92) Mean±SD	Drop-out (n=8) Mean±SD	Stay-in (n=53) Mean±SD	Drop-out (n=55) Mean±SD	Stay-in (n=40) Mean±SD	Drop-out (n=57) Mean±SD
Age	21.5± 4.4	21.8± 5.9	20.4± 3.2	22.3± 3.8	22.5± 4.6	22.7± 4.0	21.4± 3.1	22.4± 4.6	22.5± 3.6	24.6± 5.1	22.4± 3.5	22.6± 3.1
FVC	5.8± 0.7	5.8± 0.8	6.0± 0.8	5.8± 0.7	5.7± 0.8	5.9± 0.7	6.0± 0.9	5.8± 0.3	5.7± 0.6	5.5± 0.7	5.9± 0.8	5.8± 0.8
FEV_1	4.8± 0.5	4.8± 0.7	4.9± 0.6	4.8± 0.7	4.6± 0.5	4.7± 0.5	4.9± 0.7	4.6± 0.5	4.6± 0.5	4.4± 0.5	4.8± 0.7	4.7± 0.7
Positive skin test	25 (18.7%)	22 (26.5%)	35 (35.0%)	4 (30.8%)	7 (7.7%)	1 (3.9%)	28 (30.4%)	1 (12.5%)	5 (9.4%)	7 (12.7%)	9 (22.5%)	16 (28.1%)
Bronchial hyperrespon-siveness	5 (3.7%)	7 (8.4%)	5 (4.8%)	0 (0.0%)	7 (7.7%)	2 (7.7%)	11 (6.0%)	1 (12.5%)	5 (9.4%)	5 (9.1%)	2 (5.0%)	7 (12.3%)

Table 9.3 : Descriptive statistics of demographic variables and lung function values at baseline and each recall.

	Baseline		1st recall		2nd recall		3rd recall	
	Grain worker (n=217) Mean±S.D	Control (n=118) Mean±S.D	Grain worker (n=117) Mean±S.D	Control (n=100) Mean±S.D	Grain worker (n=109) Mean±S.D	Control (n=97) Mean ± S.D	Grain worker (n=53) Mean±S.D	Control (n=40) Mean±S.D
Age (yrs)	21.6±5.0	20.7 ± 3.4	22.6±4.5	21.5±3.2*	23.5 ± 4.5	22.5 ± 3.2	24.7 ± 3.7	24.6 ± 3.5
Exposure(weeks)	31.9±80.9	0.2 ± 2.0****	75.7±37.1	0.2 ± 2.2****	123.1±35.5	0.0 ± 0.0	222.1±24.9	0.0 ± 0.0
Packyears	5.5±6.2	5.5 ± 5.2	6.5±5.9	5.4 ± 4.7	7.0 ± 5.5	6.8 ± 5.1	7.4 ± 6.3	3.8 ± 3.2
FVC	5.8±0.7	5.9 ± 0.8	5.8±0.8	6.0 ± 0.8	5.6 ± 0.7	5.8 ± 0.8*	5.6 ± 0.6	5.8 ± 0.8
FEV_1	4.8 ± 0.6	4.9 ± 0.6*	4.6 ± 0.5	4.8 ± 0.7*	4.5 ± 0.5	4.7 ± 0.7*	4.5 ± 1.1	4.6 ± 1.0

*: <0.05 ; **: <0.01 ; ***: <0.001 ; **** : < 0.0001

Table 9.4 : Prevalence of symptoms and smoking status during study period

	Baseline		1st recall		2nd recall		3rd recall	
	Grain worker (n=217)	Control (n=118)	Grain worker (n=117)	Control (n=100)	Grain worker (n=109)	Control (n=97)	Grain worker (n=53)	Control (n=40)
	N (%)	N (%)	N (%)	N (%)	N (%)	N (%)	N (%)	N (%)
Wheezing	43 (19.8)	31 (26.3)	35 (29.9)	34 (34.0)	44 (40.4)	40 (41.2)	25 (47.2)	17 (42.5)
Positive Skin Test	48 (21.7)	39* (33.1)	8 (6.8)	29**** (29.0)	12 (11.0)	25** (25.8)	5 (9.4)	10* (25.0)
PC_{20}	12 (5.5)	5 (4.2)	9 (7.7)	12 (12.0)	10 (9.2)	9 (9.3)	3 (5.7)	4 (10.0)
Nonsmoker	93 (42.9)	87 (73.7)***	47 (40.2)	74 (73.3)***	47 (43.1)	72(74.2)***	22 (41.5)	34 (85.0)***
Exsmoker	19 (8.7)	4 (3.4)	13 (11.1)	8 (8.0)	16 (14.7)	8 (8.2)	6 (11.3)	3 (7.5)
Current smoker	105 (48.4)	27 (22.9)	57 (48.7)	19 (19.0)	46 (42.2)	17 (17.5)	25 (47.2)	3(7.5)

*: <0.05 ; **: <0.01 ; ***: <0.001 ; **** : < 0.0001

Table 9.5 : Survival analysis for baseline bronchial hyperresponsiveness data

	Univariate			Multivariate		
	Coefficient	Risk Ratio	95% C.I.	Coefficient	Risk Ratio	95% C.I.
Group*	0.261	1.30	0.46 - 3.68	0.355	1.43	0.47-4.37
Skin test	0.736	2.09	0.79 - 5.49	0.616	1.85	0.69-4.99
FEV_1	-0.759	0.47	0.20 - 1.10	-1.118	0.33	0.12-0.87
Smoking group	-0.223	0.80	0.20 - 1.10	-0.381	0.68	0.24-1.96
Wheezing	0.670	1.97	0.73 - 5.34	0.643	1.90	0.68-5.33
Age	-0.045	0.96	0.84 - 1.09	-0.047	0.954	0.83-1.09
Height	0.026	1.03	0.73 - 5.34	0.076	1.08	0.99-1.18

* Grain worker vs control

Table 9.6 : Logistic analysis for baseline bronchial hyperresponsiveness data

	Univariate			Multivariate		
	Coefficient	Odds Ratio	95% C.I.	Coefficient	Odds Ratio	95% C.I.
Group*	0.28	1.32	0.45 - 3.87	0.40	1.49	0.46 - 4.84
Skin test	0.75	2.12	0.78 - 5.77	0.64	1.89	0.67 - 5.39
FEV_1	-0.77	0.46	0.19 - 1.12	-1.16	0.31	0.11 - 0.89
Smoking group	-0.22	0.80	0.30 - 2.17	-0.39	0.68	0.22 - 2.05
Wheezing	0.70	2.01	0.71 - 5.64	0.68	1.97	0.66 - 5.86
Age	-0.05	0.96	0.84 - 1.09	-0.05	0.95	0.83 - 1.09
Height	0.03	1.03	0.95 - 1.11	0.08	1.08	0.99 - 1.19

* Grain worker vs control

Table 9.7 : Parameter and standard error estimates of Cox's proportional hazards model for correlated bronchial hyperresponsiveness data

	Parameter Estimate (β)	Partial likelihood S.E	WLW S.E.	Jackknife S.E.	Bootstrap S.E.
Group*	-0.1359	0.2723	0.3353	0.3485	0.3610
Skin test	0.6649	0.2681	0.3243	0.3410	0.3578
FEV_1	-1.1788	0.2681	0.3460	0.3565	0.3535
Smoking	-0.1178	0.2838	0.3429	0.3563	0.3679
Wheezing	0.6140	0.2600	0.2679	0.2759	0.2794
Age	-0.0391	0.0343	0.0360	0.03884	0.0399
Height	0.0494	0.0221	0.0237	0.02487	0.0247

* Grain worker vs control

Parzen, 1996), the estimated relative bias of partial likelihood for variances of regression estimates are given in Table 9.10. The relative bias of partial likelihood variance of regression coefficients vs. jackknife and bootstrap were similar to each other, but different from the estimated relative bias of the partial likelihood vs. the WLW method. Interactions between work status and wheezing was also tested in each of the above four mentioned models, the interaction was not significant in any of the models.

9.5 Discussion

The New Grain Workers' study was conducted because of the observation of Cotton et al. (1982) that young cereal grain workers with an average of 2.5 years of exposure to grain dust showed measurable alterations in pulmonary function tests. In the New Grain Workers' study, we obtained questionnaire information on young men commencing employment in the grain industry, within, on average, three to four months of starting work. We re-examined these grain workers and control subjects for four years. Follow-up was better in control subjects compared to grain workers. At the 2nd observation 85.6% came back for re-examination, while only 53.4% of grain workers came back. At the 3rd observation 50% of grain workers and only 17% of control subjects dropped out. At the 4th observation 24.2% of grain workers and 33.9% of control subjects came back for re-examination.

We observed a progressive increase in cough and phlegm production over the first three years (Pahwa et al., unpublished). There was an appreciable fall off in numbers of both workers and control subjects from the 3rd observation to the 4th

observation, and therefore it was difficult to interpret or draw any inferences from the data on the fourth observation, they were somewhat less reliable. DoPico et al. (1983) reported that grain dust has a potentially highly irritative effect . Our results are similar to doPico et al. (1983), the rapidity of onset of respiratory symptoms in these young grain workers following employment commencement suggests a significant irritant effect of grain dust on the respiratory epithelium. This effect on respiratory epithelium is possibly associated with increases in airways responsiveness. In another longitudinal study, subjects who had left their jobs between the first and second surveys had significantly more symptoms of cough and shortness of breath than those who were tested at both surveys (Broder et al., 1985). In our study, young grain workers had a lower prevalence of positive skin test than did control subjects at the baseline, and by the 1st recall this difference increased. As discussed by Dosman et al. (1991) skin test positivity may be a factor in job selection as well as in dropping out. This could be a factor in the healthy worker effect Dosman et al. 1991; Zejda et al., 1992).

Mink et al. (1980) demonstrated an increase in nonspecific bronchial reactivity to inhaled histamine in lifetime non-smoking grain handlers compared to lifetime non-smoking control subjects, that is not related to evidence of allergy as measured by medical history, grain dust sensitivity, prick skin test and IgE measurements. Enarson et al. (1985) studied the bronchial hyperexcitability to methacholine in grain handlers. They reported that lower levels of FEV_1, and immediate reactivity to common allergies are significant predictors of bronchial hyperexcitability. Chen (1992) studied the effect of grain dust on non-specific bronchial hyperreactivity. He studied grain elevator or mill workers divided into heavy and light exposure groups. Significant reductions in lung

function values (FEV_1, FVC, PEF, V50 and V25) were reported after provocation tests (1992). A significantly higher proportion of heavy-exposure grain workers was associated with bronchial hyperreactivity than in the light-exposure group. Chen (1992) concluded that exposure to heavy concentrations of grain dust for long periods can impair pulmonary function and cause nonspecific hyperreactivity. In our study, we observed that among those who had bronchial hyperresponsiveness, an appreciable proportion of them dropped out after baseline. Twenty three grain workers had bronchial hyperresponsiveness at baseline, of these 18 dropped out after baseline. At the 1st recall 9 grain workers had bronchial hyperresponsiveness and 7 dropped out after the 2nd recall. Among controls, those who had bronchial hyperresponsiveness, the proportion of drop-outs was comparatively less than grain workers with bronchial hyperresponsiveness.

Survival (Cox's regression model) and repeated logistic regression analysis of correlated data produced different results. Survival analysis techniques are preferable because survival analysis techniques account (i) for persons lost during the follow-up period, and (ii) for time to the occurrence of the outcome of interest. We recommend using the survival analysis with robust standard error estimates (WLW, jackknife or bootstrap) for correlated bronchial hyperresponsiveness because the partial likelihood estimates of standard errors of parameter estimates are not consistent. Lipsitz and Parzen (1996) extended the use of the jackknife from a parametric failure time model for an observation within a cluster to the semiparametric Cox's model. We extended this approach to analyze correlated bronchial hyperresponsiveness data where histamine dose was treated as survival time. In addition, bootstrap, WLW and jackknife

Table 9.8 : Risk ratios from Cox's hazard model and confidence intervals obtained by using different methods for bronchial hyperresponsiveness data.

	Risk ratio	.Partial likelihood 95% C.I.	WLW 95% C.I.	Jackknife 95% C.I	Bootstrap 95% C.I.
Group	0.87	0.51 - 1.49	0.45 - 1.68	0.44 - 1.73	0.43 - 1.77
Skin test	1.94	1.15 - 3.29	1.03 - 3.67	1.00 - 3.79	0.96 - 3.92
FEV_1	0.31	0.18 - 0.52	0.16 - 0.61	0.15 - 0.62	0.15 - 0.62
Smoking	0.89	0.19 - 1.55	0.17 - 0.64	0.16 - 0.66	0.16 - 0.67
Wheezing	1.85	1.11 - 3.08	1.09 - 3.12	1.08 - 3.17	1.07 - 3.20
Age	0.96	0.90 - 1.03	0.90 - 1.03	0.89 - 1.04	0.89 - 1.04
Height	1.05	1.01 - 1.10	1.00 - 1.10	1.00 - 1.10	1.00 - 1.10

Table 9.9 : Repeated logistic analysis for correlated hyperresponsive data

	Parameter Estimate (β)	Robust S.E.	Odds ratio	95% C.I.
Group	0.04	0.401	1.04	0.48 - 2.29
Skin test	0.80	0.333	2.22	0.48 - 4.25
FEV_1	-1.64	0.365	0.19	0.09 - 0.40
Smoking	-0.26	0.395	0.77	0.36 - 1.67
Wheezing	0.31	0.287	1.36	0.78 - 2.39
Age	-0.04	0.031	0.96	0.91 - 1.02
Height	0.08	0.026	1.09	1.04 - 1.14

Table 9.10 : Estimated relative bias of the partial likelihood estimates of variance of regression coefficients [Var(β_i)]

	Partial likelihood Var(β_i)	WLW Var(β_i)	Jackknife Var(β_i)	Bootstrap Var(β_i)	Estimated relative bias		
					WLW vs.[@] partial likelihood	Jackknife vs [#] partial likelihood	Bootstrap vs[$] partial likelihood
Group	0.7412	0.1124	0.1214	0.1303	34.02	38.92	43.09
Skin test	0.0719	0.1052	0.1163	0.1280	31.69	38.21	43.86
FEV1	0.0719	0.1197	0.1271	0.1250	39.99	43.46	42.50
Smoking group	0.0806	0.1176	0.1270	0.1353	31.50	36.55	40.49
Wheezing	0.0676	0.0718	0.0762	0.0781	5.80	11.21	13.40
Age	0.0012	0.0013	0.0015	0.0016	9.31	21.95	26.26
Height	0.0005	0.0006	0.0006	0.0006	12.76	20.74	20.10

@ : relative bias of WLW vs. partial likelihood = [(Var(WLW) - Var(partial likelihood))/Var(WLW)] * 100

#: relative bias of jackknife vs. partial likelihood = [(Var(jackknife) - Var(partial likelihood))/Var(jackknife)] * 100

$: relative bias of bootstrap vs. partial likelihood = [(Var(bootstrap) - Var(partial likelihood))/Var(bootstrap)] * 100

methods were used to obtain the estimates of the variance. In the bootstrap technique, subjects were treated as the resampling units. In general, the jackknife estimator performs less well than the bootstrap, and requires less computation (Efron, 1982). In our study, the results of jackknife and bootstrap analyses were very comparable.

In any longitudinal data analysis, missing values arise whenever the sequence of measurements/observations from subjects within the study is incomplete. This incompleteness can be due to one of the following reasons: (i) the intended measurements/observations were not taken; or (ii) loss to follow up or drop outs. In our study, the missing observations were due to subjects dropping out from the study. Although the exact reasons for subjects dropping out are not known, no systematic differences were found between stay-in and drop-out subjects

In conclusion, correlated survival analysis is a useful technique to analyze bronchial hyperresponsiveness data. Consistent estimates of the variance of regression parameters can be obtained by jackknife and bootstrap techniques by writing a simple macro. With larger data sets, the jackknife method is preferred over the bootstrap method, because one can expect computational savings.

10. General Summary

10.1 Introduction

In this thesis, i) different statistical models were explored to predict longitudinal changes in lung function measurements in grain elevator workers; ii) recently developed goodness of fit statistics were applied to assess the fit of these models; iii) survival data techniques were used to determine the predictors for first episode of wheezing; and iv) censored survival data techniques for correlated data were used to determine the predictors for bronchial hyperresponsiveness.

10.2 Statistical models to predict longitudinal changes in lung function measurements:

Marginal, transitional and random effects models were fitted using four different covariance structures: i) independence/uncorrelated; ii) compound symmetric; iii) unspecified and iv) autoregressive assuming equally spaced observations. We also fitted a random effects model with autoregressive covariance structure assuming unequally spaced observations.

Marginal and transitional models were fitted by using generalized estimating equations (GEE) approach, a general approach introduced by Liang and Zeger (1986) and Zeger and Liang (1986). The random effects models were fitted by using restricted maximum likelihood method (Patterson and Thompson, 1971). Random effects models

have their advantages, but at the same time, the estimation of the parameters of a random-effects linear model by iterative maximum likelihood methods/restrictive maximum likelihood models can be computationally problematic, especially when the data set is quite large (Ware, 1985). We encountered numerical problems while fitting random effects models and were unable to fit these models for some of the specified covariance structures.

The transitional model with independence covariance structure was used to estimate the annual decline in pulmonary function test values in different age categories, and different exposure and smoking categories before and after dust control. Rosner et al. (1985) were the first to apply autoregressive models (previous lung function as one of the covariables) to analyze lung function data. Ware et al. (1989) used this approach to analyze lung function data from six cities study. Pahwa et al. (1994) used this approach to analyze data from first three cycles of Labour Canada GDMSP. It was observed that if previous lung function was one of the covariables in the model, specification of covariance structure had little effect on the results and gave a better fit compared to marginal and random effects models. Before dust control: i) the yearly loss in FEV_1 increased with increasing years in the grain industry for ex-smoking grain workers, but not for non-smokers and current smokers; ii) among the subjects who were in the grain industry for less than 10 years, the yearly loss in FEV_1 was greatest among the ex-smoking grain workers (-43.4 ml/yr), but there was no difference in the yearly losses of non-smoking (-36.5 ml/yr) and current smoking (-36.3 ml/yr) grain workers; and iii) Ex-smokers and current smokers who were in the industry for more than 20 years had greater declines in FEV_1 than did non-smokers. After dust

control: i) among the subjects who were in the industry for less than 10 years, the yearly loss in FEV_1 was greatest in the current smokers (-9.0 ml/yr); ii) the yearly loss for ex-smokers (-7.5 ml/yr) and current smokers (15.0 ml/yr) who were in the industry for 10 to 20 years had greater yearly loss in FEV_1 than did non-smokers; and iii) among workers with 20 years or more in the industry, the yearly loss was greater in ex-smokers (-15.3 ml/yr) and current smokers (-9.1 ml/yr) than among non-smokers. After dust control: i) the yearly loss in FVC was greatest for ex-smoking workers who were in the industry for less than 10 years; ii) the non-smokers who were in the industry for 10 to 20 yr. had greatest annual decline (-139.3 ml/yr) than did ex-smokers and current smokers; and iii) in workers with 20 yr or more in the grain industry, the yearly loss in FVC was similar in non-smokers, ex-smokers, and current smokers. Reduction in the yearly losses show that dust control measurements definitely had a positive affect on the respiratory health of grain workers.

An important finding from the analysis of the longitudinal lung function data was that REM models with unequally-spaced AR(1) covariance structure and REM models with unequally-spaced AR(1) with observational error are the most appropriate models for prediction of FEV_1 and FVC respectively. The estimated observational error (=0.3870) for forced vital capacity (FVC) was higher than the estimate of observational error (=0.0822) for forced expiratory volume in the first second (FEV_1). These results support the clinical observation that higher observational error is associated with FVC measurements compared to FEV_1 measurements.

10.3 GOODNESS OF FIT

In this thesis, recently developed goodness-of-fit statistics (Vonesh et al., 1996) were used to assess the goodness-of-fit of marginal and transitional models fitted by using GEE approach, and REM fitted by restrictive likelihood methods. These new goodness-of-fit were, the concordance correlation coefficient, r_c to test the adequacy of the response model was used, the concordance correlation coefficient, $r(\hat{\omega})$, to study the closeness between $\hat{\Omega}_R$ and $\hat{\Omega}$, and the pseudo-likelihood ratio test, $\hat{\lambda}$ for testing the null hypothesis that the assumed (specified) and true covariance structures are equal. Assuming data were equally spaced, and based on these goodness-of-fit statistics, it was concluded that transitional models for FEV_1 and FVC gave better fit compared to marginal and random effects models. The values of r_c were close to 1 for transitional model for FEV_1 (ranging from 0.9297 to 0.9298 for different covariance structures, when age was in the model; and from 0.9286 to 0.9292 for different covariance structures when age was not in the model) and this could be due to the inclusion of previous FEV_1 as one of the covariates in the model. There are some limitations to the statistics $\hat{\Omega}_R$ and $\hat{\Omega}$. The concordance correlation coefficient $\hat{\Omega}_R$, to measure the closeness between specified and true covariance structures had very limited range, as it was seen for all the three types of models fitted assuming different covariance structures: (i) the range of $\hat{\Omega}_R$ was 0.8003 - 0.9501 for marginal models with age in the model and 0.7978 - 0.9440 for marginal models with age not in the model; (ii) the range of $\hat{\Omega}_R$ was 0.9280 - 0.9355 for transitional models with age in the model and 0.9297 - 0.9361 for transitional models with age not in the model; and (iii) the range of $\hat{\Omega}_R$ was 0.9249 - 0.9643 for random effects model with age in the model and 0.9314 -

0.9632 for random effects models with age not in the model. Similar results were obtained for FVC.

The concordance correlation coefficient has certain advantages and disadvantages. There are two advantages of the correlation concordance coefficient: (i) one may not need to fit a family of nested models and this may help save time at the model building stage; and (ii) the concordance correlation coefficient used to assess the adequacy of response function is semi-nonparametric, because it does not require specification of a likelihood function. A major disadvantage of these goodness-of-fit statistics are that they are not suitable for discrete data. The values of pseudo-likelihood ratio test, $\hat{\Omega}$ were very high for GDMSP data, because $\hat{\Omega}$ is a function of sample size and increase with increasing n (sample size). All new goodness-of-fit statistics developed by Vonesh et al. (1996) to assess the adequacy of longitudinal models and variance-covariance structures need further experience to determine the full utility of these methods.

10.4 Application of survival analysis technique to longitudinal data

The Cox proportional hazard model was used to determine the predictors of first episode of wheeze among initially symptom-free Canadian grain elevator workers studied longitudinally (GDMSP). Based on the analysis, the predictors of first episode of wheezing were: current smoking status; and FEV_1/FVC ratio at the baseline. Several conclusions were drawn from this analysis of longitudinal data of grain workers who participated in four different surveillance cycles over a 12 year period. Grain workers who had baseline FEV_1/FVC ratio < 70% were at risk of developing wheeze. Current

smokers had a relative risk of 2.3 of developing wheeze compared to nonsmokers when adjusted for age, height, exsmoking status and FEV_1/FVC ratio. The group of nonsmokers with FEV_1/FVC ratio $\geq 70\%$ had the best survival function for remaining wheeze free as compared to i) nonsmokers with $FEV/FVC < 70\%$; ii) smokers with $FEV_1/FVC < 70\%$; and iii) smokers with $FEV_1/FVC < 70\%$. The risk of new onset of wheezing also increased with age.

10.5 Models for correlated survival data

Survival analysis techniques for correlated data were used to determine the predictors of airway responsiveness. In this approach, the dose of the stimulus (histamine) was considered to be time to failure and PC_{20} as outcome. Subjects who did not reach a 20% fall in FEV_1 after the maximum dose were considered censored. Using survival analysis techniques, full use is made of the information that some PC_{20} values were known to exceed the maximum dose, instead of assuming them to be equal to the maximum dose as in traditional analyses.

The regression estimate was obtained, by assuming that observations were independent. The regression estimator $\hat{\underline{\beta}}$ obtained by the using partial likelihood (Cox, 1972) is consistent. However the variance-covariance matrix estimator for $\hat{\underline{\beta}}$ may no longer be valid for inferential purposes. So, a valid covariance matrix estimate which accounted for the dependence among related observations was obtained by three different methods, WLW, jackknife, and bootstrap. The regression estimate $\hat{\underline{\beta}}$ was obtained by using PROC PHREG (SAS technical report, P-229, 1992), but the estimation of the covariance matrix needed extra programming. The relative efficiency of these robust estimates were compared with covariance matrix obtained by partial

likelihood method, naively assuming that observations on each subject were independent.

There was not much difference between the results of the jackknife and bootstrap methods. The significant predictors of bronchial hyperresponsiveness were : a positive allergy skin test, low FEV_1, smoking status and wheezing.

10.6 Recommendations for future analysis

10.6.1 Missing observations

Missing values arise whenever one or more observations or a sequence of observations from subjects within the study are incomplete because of several reasons: i) intended measurements were not taken, ii) were lost; or iii) otherwise unavailable. It was explained in the first few chapters that some statistical methods can be used to analyze data from well balanced designs in which measurements were collected at a common set of points on all units/subjects. Missing values result in unbalanced data and raise some analytical difficulties. It is very important to know why the values are missing and if these missing values had any effect on the practical question or research question that we want to answer by analyzing the available data, e.g. if the subjects non-participation is related to study outcome, then it is important to model that.

Diggle et al (1995) explained the above point with the help of an example for non-longitudinal data. We can face similar situations in longitudinal studies. Diggle et al (1995) gave some details about how to test the completely random drop-outs and how to model the random drop-out process. Little and Rubin (1987), gave a general treatment of statistical analysis with missing values, which includes a useful hierarchy of missing data.

In longitudinal studies an important distinction should be made by determining whether missing values occur intermittently/irregularly or as drop-outs (Diggle et al, 1995). Suppose we have an array of measurements $Y_1, Y_2, \ldots, Y_k, \ldots, Y_n$ on a particular subject. We say missing values occur as drop-outs if whenever Y_j is missing, so are Y_k for all $k \geq j$, otherwise we say missing values are intermittent. Drop-outs arise because some subjects withdraw from the study prematurely. Drop-outs do not arise as a result of censoring applied to individual measurements (Diggle et al, 1995). Subjects dropout from studies for reasons directly or indirectly related to the measurement process. In clinical trials subjects/units may have to drop out from the study because of ethical reasons. For example, in long-term trials of drugs to reduce blood pressure, it is unethical to continue testing a drug on patients if the patients' blood pressure is not adequately controlled (Murray and Findlay, 1988). So it is important to distinguish amongst completely random, random and informative dropouts and to use the proper methods to analyze data by using this additional structure of dropout data.

When intermittent missing values arise from a mechanism unrelated to the measurement process, they can be assumed to be completely random and the resulting data can be analyzed by any methods which can handle unbalanced data. Robins et al (1995) proposed methods which can handle data that is missing at random. The methods proposed by Robins et al (1995) work well only for dropouts. In longitudinal studies, often subjects miss a single observations and then are seen again, e.g. in GDMSP, this situation occurred. Lee, Laird, and Johnston (in revision) modified the GEE approach that combines REML estimating equations for the parameters in the variance-covariance matrix. This new approach considers not only the model for the

mean, but the model for the variance and the underlying missingness process (Horton and Lipsitz ,1999).

In out Grain Dust Medical Surveillance Program (GDMSP) we had dropouts, intermittent missing values and we also had several missing observations for the variable. "years in the grain industry" at Cycle I (Oct. 1978 - Sept. 1981). In utilizing the GEE methodological approach to fit marginal and transitional models and the maximum likelihood approach to fit random effect models, the subjects for whom this information was missing, were not included in the analysis. Pahwa et al (1994) extrapolated these values using the information available at the next two cycles (Cycle II and Cycle III). A large number of grain elevator workers dropped out from the study and we do not know the exact reasons for dropouts. If these dropouts were not at random, the next step could be to correct for bias caused by dropouts in GDMSP.

10.6.2. MEASUREMENT ERROR MODELS

In epidemiological studies, one is often forced to use surrogate variables in place of factors which cannot be measured directly. How this surrogate information effects the regression estimation will be an interesting area of research to explore.

When the exact values of covariates in Cox's proportional hazards model are not observed and only surrogate covariates are available, the usual partial likelihood estimates of the surrogate covariates are asymptotically biased and underestimate the risk. Several methods have been proposed for correcting the partial likelihood for measurement error in Cox's proportional hazards model (Hughes 1993; Nakamura 1992, and Nakamura and Akazawa 1994). In the Grain Dust Medical Surveillance Program,

we used length of time spent working in various locations in a grain elevator as a surrogate of the cumulative exposure to grain dust. This surrogate information is not a direct measure of dust level. In the grain workers study, the grain workers have different levels of exposure to grain dust, but we used years in the industry as an exposure measurement for all of them, which is not an accurate indicator of exposure to grain dust. It is a common practice to ignore the potential lack of representativeness of these surrogate variables and to use them directly as observed in multivariable modelling. It is of interest to examine the possible effect of such measurement errors on estimated relative risks due to grain dust, and to use this information to develop improved risk estimators. In particular, ignoring measurement error can lead to the wrong conclusions, for example the true but unobserved data may indicate a positive effect, while the observed data indicate the opposite. Synonyms for 'predictors measured with error' are 'errors-in-variables' and 'measurement-error models' (Caroll, 1989). By these phrases we mean that there is a true value of a predictor, but we can observe only an estimate of this true value.

In a recent modeling approach, Wulfsohn and Tsiatis (1997) developed a joint model for survival and longitudinal data measured with error. In this methodology a joint maximization of the likelihood from both the covariate process and the survival data occurs. In this approach a two-stage model was fitted. In the first stage the covariate is modeled using growth curve models with random effects (Laird and Ware, 1982). In the second stage, the modeled value is substituted into the partial likelihood for the Cox model with time-dependent covariates, and the partial likelihood is then maximized (Dafni, 1993; Tsiatis, DeGruttola, and Wulfsohn, 1995). This approach has

been supported on the basis that it reduces the bias of the parameter estimate in Cox models.

In occupational and environmental studies, exposure typically cannot be measured exactly and ignoring this measurement error leads to asymptotically biased estimators of the threshold. For GDMSP data, a proportional hazard model was used to determine the predictors of time to first wheezing episode in the surveillance program. In this analysis, there is a need to investigate how surrogate information on grain dust measurements affects the estimates of predictors of time to first wheeze in the proportional hazards model. The approach developed by Wulfsohn and Tsiatis (1997) can be applied to correct the bias caused by using surrogate information in the analysis of GDMSP data.

11. Bibliography

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In: *2nd International Symposium on Information Theory and Control*, Eds., Petrov, E.B.N., and Csaki, F., Budapest: Akademia Kiado, pp. 267 - 281

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19 : 716 - 723.

Anderson, T. W. (1984). *An introduction to multivariate statistical analysis.* Wiley, New York.

Bahadur, R.R. (1961). A representation of the joint distribution of responses to n dichotomous items. In: *Studies on item analysis and prediction.* Eds., Solmon, H., Stanford Mathematical Studies in Social Sciences VI, Stanford University Press, Stanford, California, pp. 158 - 168.

Barnhart, H. X., and Williamson, J.M. (1998). Goodness-of-fit tests for GEE modelling with binary responses. *Biometrics*, 54 : 720 - 729.

Bishop, Y.M.M., Fienberg, S.E., and Holland, P.W. (1975). *Discrete multivariate analysis: theory and practice.* MIT Press, Cambridge, Massachussets.

Box, G.E.P. (1950). Problems in the analysis of growth and wear curves. *Biometrics*, 6: 362 - 389.

Broder, I., Corey, P., Davies, G., Hutcheon, M., Mintz, S., Inouye, T., Hyland R., Leznoff, A., and Thomas, P. (1985). Longitudinal study of grain elevator and control workers with demonstration of healthy worker effect. *Journal of Occupational Medicine*, 27: 873 - 880

Burney, P.G., Chinn, S., Britton, J.R., Tattersfield, A.E., Papacosta, A.O. (1989). What symptoms predict bronchial response to histamine? Evaluation in a community survey of the bronchial symptoms questionnaire (1984) of the International Union Against Tuberculosis and Lung Disease. *International Journal of Epidemiology*, 18 : 167-173.

Carroll, R.J., Speigelman, C.H., Lan, K.K.G., Bailey, K.T. and Abbott, R.D. (1989). On error in variables for binary regression models. *Biometrika*, 71 : 19 - 25.

Carter, R.L, Reanick, M.B., Ariet, M., Shieh, G., and Vonesh, E.F. (1992). A random coefficient growth curve analysis of mental development in low-birth-weight infants. *Statistics in Medicine*, 11: 243 - 256.

Chambers, J.M., Cleveland, W.S., Kleiner, B. and Tukey, P.A. (1983). *Graphical methods for data analysis.* Belmont, CA, Woodworth, International group.

Chan-Yeung, M., Wong, R., MacLean, L. (1979). Respiratory abnormalities among grain elevator workers. *Chest*, 75 : 461-467.
Chan-Yeung, M. (1990). Grain dust asthma, does it exist? In: *Principles of Health and Safety in Agriculture*, Eds., Dosman, J.A., Cockcroft, D.W., CRC Press, Boca Raton, Florida, 169 - 174.

Chan-Yeung, M., Schulzer, M., Maclean, L., Dorken, E., Tan, F., Lam, S., Enarson, D., Grzybowski, S. (1981). A follow-up study of the grain elevator workers in the Port of Vancouver. *Archives of Environmental Health*, 36: 75 - 81.

Chen, P. (1992). The effect of grain dust on non-specific bronchial hyperreactivity. Chinese journal of Tuberculosis, 15 : 28-30.

Christiani, D.C., Ye, T.T., Zhang, S., Wegman, D.H., Eisen, E.A., Ryan, L.A., Olenchock, S.A., Pothier, L., Dai, H.L. (1999). Cotton dust and endotoxin exposure and long-term decline in lung function: results of a longitudinal study. American *Journal of Industrial Medicine*, 35 : 321-331.

Cline, M.G., Dodge, R., Lebowitz, M.D., and Burrows, B. (1994). Determinants of percent predicted FEV_1 in current asthmatic subjects. *Chest*, 106 : 1089 - 1093.

Cockroft, D.W., Killian, D.N., Mellon, J.J., Hargreave, F.E. (1977). Bronchial reactivity to inhaled histamine: a method and a clinical survey. *Journal of Allergy and Clinical Immunology*, 7 : 235 - 243.

Cole, J.W.L., and Grizzle, J.E. (1966). Applications of multivariate analysis of variance to repeated measurements experiments. *Biometrics*, 22 : 810 - 828.

Colton, T. (1974). *Statistics in Medicine*. Little Brown and Company, Boston.

Cotton D.J., Graham, L., Li, K.Y.R., Froh, F. Barnett, G.D., Dosman, J.A. (1982). Effects of smoking and occupational exposure on peripheral airway function in young cereal grain airways. *American Review of Respiratory Diseases*, 125 : 660-665.

Cox, D.R. (1972). Regression models and life tables. *Journal of the Royal Statistical Society, Series B*, 24 : 187 - 220.

Dafini, U. (1993). *Evaluating surrogate markers of clinical outcome when measured with error*. Ph.D dissertation, Biostatistics Library, Harvard School of Public Health.

Dales, R.E., Ernst, P., Hanley, J.A., Battista, R.N., Becklake, M.R. (1987). Prediction of airway reactivity to a standardized respiratory symptom questionnaire. *American Review of Respiratory Diseases*, 135 : 817-821

Dawson, K.S., Gennings, C., and Carter, W.H. (1997). Two graphical techniques useful in deleting correlation structure in repeated measures data. *Journal of the American Statistical Association*, 51 : 275 - 283.

Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B*, 39 : 1 - 38.

Diem, J.E., and Liukkonen, J.R. (1988). A comparative study of three methods for analysing longitudinal pulmonary function data. *Statistics in Medicine*, 7 : 19 - 28.

Diggle, P.J. (1988). An approach to the analysis of repeated measurements. *Biometrics*, 40 : 959 - 971.

Diggle, P., Lang, K., Zeger, S. (1995). *Analysis of Longitudinal Data.* Oxford Science Publication.

Dimich-Ward H.D., Kennedy, S.M., Dittrick, M.A., DyBuncio, A., Chan-Yeung, M. (1995). Evaluation of the respiratory health of dock workers who load grain cargoes in British Columbia. *Occupational Environmental Medicine*, 52 : 273-278.

doPico, G.A., Jacobs, S., Flaherty, D., Rankin, J. (1982). Pulmonary reaction to durum wheat. A constituent of grain dust. *Chest*, 81 : 55-61.

Dorf, R.C. (1969). *Matrix algebra - a programmed introduction.* John Wiley and Sons Inc, New York.

Dosman JA, McDuffie HH, Pahwa P, Hall D. (1987). Statistical analysis of the environmental and medical surveillance programme in the grain industry. Labour Canada Report.

Dosman, J.A., McDuffie, H.H., Pahwa, P. (1991). Atopic status as a factor in job decision making in grain workers. *Journal of Occupational Medicine*, 33(9) : 1007-1010.

Dwyer, J.H. (1992). *Statistical models for longitudinal studies of health.* Monographs in Epidemiology and Biostatistics. V. 16. Oxford Univ Press, New York.

Efron, B. (1982). *The Jackknife, the Bootstrap, and other Resampling Plans.* CBMS-NSF regional conference series in applied Mathematics, 38. SIAM monograph, Philadelphia, PA.

Efron, B., and Tibshirani, R.J. (1993). *An Introduction to the Bootstrap.* Chapman and Hill, New York.

Efron, B., and Lepage, R. (1992). Introduction to bootstrap. In: *Exploring the Limits of Bootstrap*, Eds., Lepage R., and Billard, L., Wiley, New York, pp. 3 - 10.

Elston, R.C. (1964). On estimating time response curves. *Biometrics*, 20, 643 - 647.

Elston, R.C., and Grizzle, J.E. (1969). Estimation of time-response curves and their confidence bands. *Biometrics*, 18 : 148 - 159.

Enarson, D.A., Vedal, S., Chan-Yeung, M. (1985). Rapid decline in FEV1 in grain handlers. *American Review of Respiratory Diseases*, 132 : 814-817.

Feng, Z., McLerran, D., and Grizzle, J.E. (1996). A comparison of statistical methods for clustered data analysis with gaussian error. *Statistics in Medicine*, 15 : 1793 -1806.

Ferretti, B., Refini, R.M., Pieroni, M.G., Sestini, P., and Vagliasindi, M. (1993). The methacholine challenge test in the diagnosis of respiratory diseases. *European Respiratory Journal*, 6 (Suppl 17) : p1427 - 473s.

Ferris, B.G. Epidemiology stadardization project-II. (1978). Recommended respiratory disease questionnaires for use with adults and children in epidemiological research. *American Review of Respiratory Diseases*, 118 : 7 - 53.

Fitzmaurice, G.M., Laird, N.M., and Rotnitzky A.G. (1993). Regression models for discrete longitudinal responses. *Statistical Science*. 8 : 284 - 309.

Glencross, P.M., Weinberg, J.M., Ibrahim, J.G., Christiani, D.C. (1997). Loss of lung function among sheet metal workers: ten-year study. *American Journal of Industrial Medicine*, 32(5) : 460-466.

Glindmeyer, H.W., Lefante, J.J., Jones, R.N., Rando, R.J., Kader H.M.A., and Weill H. (1991). Exposure-related declines in the lung function of cotton textile workers. *American Review of Respiratory Diseases*, 144 : 17 - 22.

Gottlieb, D.J., Sparrow, D., O'Connor, G., and Weiss, S. (1996). Skin test reactivity to common aeroallergens and decline of lung function: The normative aging study. *American Journal of Respiratory and Critical Care Medicine*, 153 : 561 - 566.

Grady, J.J., and Helms, R.W. (1995). Model selection techniques for the covariance matrix for incomplete longitudinal data. *Statistics in Medicine*, 14 : 1397 - 1416.

Grizzle, J.E., and Allen D.M. (1969). Analysis of growth and dose response curves. *Biometrics*, 25 : 357 - 382.

Guidelines for an Environmental and Medical Surveillance Program in the Grain Industry. (1978). "file 897-7-11", OSHB, Labour Canada.

Hartley, H.O., and Rao, J.N.K. (1967). Maximum likelihood estimation for the mixed analysis of variance model. *Biometrika*, 54 : 93 - 108.

Harville, D.A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association*, 72 : 320 - 340.

Healy, M.J.R. (1961). Experiments for comparing growth curves. Abstract No. 752. *Biometrics*, 17 : 333.

Henderson, C.R. (1953). Estimation of variance and covariance components. *Biometrics*, 9 : 226 - 252.

Henderson, C.R. (1984). *Applications of linear models in Animal Breeding*. University of Guelph, Guelph, Ont.

Horton, N.J., and Lipsitz, S.R. (1999). Review of software to fit generalized estimation equations regression models. *The American Statistician*, 53 : 160 - 169.

Hughes, M.D. (1993). Regression dilution in the proportional hazards model. *Biometrics*, 49 : 1056 - 1066.

Huy, T., De Schipper, K., Chan-Yeung, M., Kennedy, S.M. (1991). Grain dust and lung function. Dose-response relationships. *American Review of Respiratory Diseases*, 144 : 1314-1321.

Jaakkola, M.S., Jaakkola, J.J.K., Ernst, P., and Becklake, M.R. (1993). Respiratory symptoms in young adults should not be overlooked. *American Review of Respiratory Diseases*, 147 : 359 - 366.

Jedrychowski, W., Krzyanowski, M., Wyoski, M. (1988). Are chronic wheezing and asthma-like attacks related to FEV1 decline? The Cracow Study. *European Journal of Epidemiology*, 4 : 335 - 342.

Jin, S., and Sherrill, D.L. (unpublished, 1997). Comparison of methods for analysing longitudinal pulmonary function data using Monte Carlo Simulations. Tuson, Arizona,

Jones, R.H. (1987). Time series analysis with unequally spaced data. In: Hannan, E.J., Krishnaiah, P.R., Rao, M.M., eds. *Handbook of Statistics*, vol. 5: time series in the time domain. Amsterdam: North-Holland Publishing Company, 157-177.

Karim, M. and Zeger, S. L. (1988). GEE. A SAS Macro for longitudinal data analysis. *Technical Report*, Department of Biostatistics, The Johns Hopkins University, Baltimore, MD.

Kongerud, J., and Samuelsen, S.O. (1991). A longitudinal study of respiratory symptoms in aluminium potroom workers. *American Review of Respiratory Diseases*, 144 : 10 - 16.

Krzyanowski, M., Sherrill, D.L., and Lebowitz, M.D. (1990). Longitudinal analysis of the effects of acute lower respiratory illnesses on pulmonary function in an adult population. *American Journal of Epidemiology*, 131 : 412 - 422.

Krzyzanowski, M., Camilli, A.E., and Lebowitz, M.D. (1990). Relationships between pulmonary function and changes in chronic respiratory symptoms. Comparison of Tucson and Cracow longitudinal studies. *Chest*, 98 : 62 - 70.

Kundson, R.J., Lebowitz, M.D., Holberg, C.J., and Burrows, B. (1983). changes in the maximum expiratory flow volume curve with growth and aging. *American Review of Respiratory Diseases*, 127 : 725 - 734.

Kunsch, H. (1989). The jackknife and the bootstrap for general stationary observations. *Annals of Statistics*, 17 : 1217 - 1241.

Kvalseth, T.O. (1985). Cautionary note about R^2. *The American Statistician*, 39 : 279 - 285.

Laird, N. M., and Ware, J. H. (1982). Random effects models for longitudinal data. *Biometrics*, 38 : 963 - 974.

Lebowitz, M.D., Knudson, R.J., Burrows, B. (1975). The Tucson epidemiological study of obstructive lung disease: I. Methodology and prevalence of disease. *American Journal of Epidemiology*, 102 : 137 - 152.

Lee, H., Laird, N.M., and Johnston, G. (in revision-1999). Combining GEE and REML for estimation of generalized linear model with incomplete multivariate data.

Lee, E.T. (1992). *Statistical Methods for Survival Data Analysis.* John Wiley and Sons, New York..

Lee, E.W., Wei, L.J., and Amato, D.A. (1992). Cox-type regression analysis for large numbers of small groups of correlated failure time observations. In: *Survival analysis: State of Art*, Eds., Klein J.P., and Goel P.K., Kluwer academic publishers, Neterlands, pp. 237 - 247.

Leech, F.B., and Healy, M. J. R. (1959). The analysis of experiments on growth rate. *Biometrics*, 15 : 98 - 106.

Liang, K-Y, and Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73 : 13 - 22.

Liang, K-Y, Zeger, S.L., and Qaqish, B. (1992). Multivariate regression analyses for categorical data (with discussion). *Journal of the Royal Statistical Society, Series B*, 54 : 3 - 40.

Lin, L.I-K. (1989). A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, 45 : 255 - 268.

Lindstorm, M.I., and Bates, D.M. (1988). Newton-Raphson and EM algorithm for linear mixed-effects models for repeated-measures data. *Journal of the American Statistical Association*, 83 : 1014-1022.

Lipsitz, S.R., Dear, K.B.G., and Zhao, L. (1994). Jackknife estimators of variance for parameter estimates from estimating with applications to clustered survival data. *Biometrics*, 50 : 842 - 846.

Lipsitz, S.R., and Parzen, M. (1996). A jackknife estimator of variance for Cox regression for correlated survival data. *Biometrics*, 52 : 291 - 298.

Little, R.J.A., and Rubin, D.B. (1987). *Statistical analysis with missing data*. John Wiley, New York.

Longford, N. T. (1993). Random coefficients models. Oxford University Press, New York.

Manor, O., and Kark, J.D. (1996). A comparative study of four methods for analysing repeated measures data. *Statistics in Medicine*, 15 : 1143 - 1159.

McCullagh P., and Nelder, J. (1989). *Generalized Linear Models*. 2nd edn, Chapman and Hall, London.

McDuffie HH, Dosman JA, Pahwa P, To T. (1989). Final report, Statistical analyses of the environmental and medical surveillance program in the grain industry. Cycle III (1984-1987). Labour Canada Report.

McDuffie, H.H., Pahwa, P., and Dosman, J.A. (1992). Respiratory health status of 3098 Canadian grain workers studied longitudinally. *American Journal of Industrial Medicine*, 20 : 753 - 762.

Mink, J.T., Gerrard, J.W., Cockcroft, D.W., Cotton, D.J., Dosman, J.A. Increased bronchial reactivity to inhaled histamine in non-smoking grain workers with normal lung function. *Chest*, 77 : 28-31.

Murray, G.D., and Findlay, J.G. (1988). Correcting for the bias caused by drop-outs in hypertension trials. *Statistics in Medicine*, 7 : 941 - 946.

Nakamura, T. (1992). Proportional Hazards model with covariates subject to measurement error. *Biometrics*, 48 : 829 - 838.

Nakamura, T., and Akazawa, K. (1994). Corrected likelihood for proportional hazards measurement error model and its application. *Environmental Health Perspectives*, 102 (Suppl 8) : 21 - 24.

Pahwa, P., Hall, D., McDuffie, H.H, Dosman, J.A. (unpublished, 1991). A prospective evaluation of respiratory ocular and dermatological symptoms in grain workers from employment commencement to four years' exposure.

Pahwa P., Wang P., Senthilselvan A., Dosman J.A., McDuffie H.H. (1994). Statistical analyses of the environmental and surveillance program in the grain industry. Cycle IV (1987-1990), Labour Canada Report.

Pahwa, P., Senthilselvan, A., McDuffie, H.H., and Dosman, J.A. (1994). Longitudinal estimates of pulmonary function decline in grain workers. *American Journal of Respiratory and Critical Care Medicine*, 150 : 656 - 662.

Pahwa, P., Senthilselvan, A., McDuffie, H.H., Dosman, J.A. (1998). Predictors of onset of wheezing in grain elevator workers. *Canadian Respiratory Journal*, 5 : 200-205.

Patterson, H.D., and Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika*, 58 : 545 - 554.

Potthoff, R.F., and Roy, S.N. (1964). A generalized multivariate analysis of variance model useful especially for growth curve problems. *Biometrika*, 51 : 313 - 326.

Prentice, R.L., and Zhao, L.P. (1991). Estimating equations for parameters in Means and covariance of multivariate discrete and continuous responses. *Biometrics*, 47 : 825 - 839.

PROC MIXED - *SAS technical report p-229, changes and enhancement, Release 6.07*,SAS institute Inc. SAS campus Drive, Cary, NC 27513.

Rao, C.R. (1958). Some statistical methods for comparison of growth curves. *Biometrics*, 14 : 1 - 17.

Rao, C.R. (1959). Some problems involving linear hypotheses in multivariate analyses. *Biometrika*, 46: 49 - 58.

Rao, C.R. (1961). Some observations on multivariate statistical methods in anthropological research. *Bulletin of International Statistical Institute*, 38 : 99 - 109.

Rao, C.R. (1965). The theory of least squares when the parameters are stochastic and its application to the analysis of growth curves. *Biometrika*, 52 : 447 - 458.

Rao, C.R., and Mitra, S.K. (1971). *Generalized inverse of matrices and its applications.* John Wiley and Sons Inc., New York.

Rao, C.R. (1973). *Linear statistical inference and its application.* Second edition. John Wiley and Sons, New york.

Rijcken, B., Schouten, eiss, S., Segal, M., Speizer, Van Der Lende R. (1988). Longitudinal analysis of the relationship between bronchial hyperactivity and pulmonary function. In: Sluiter, H., Van der Lende R, eds. Proceedings of the Fourth International Symposium on Bronchitis. Bronchitis IV. Assen. The Netherlands: Van Gorcum, 94-106.

Robins, J.M., Rotnizky, A., and Zhao, L.P. (1995). Analysis of semi-parametric regression methods for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association*, 90 : 106 - 121.

Rosner, B., Munoz, A., Tager, I., Speizer, F., and Weiss, S. (1985). The use of an autoregressive model for the analysis of longitudinal data in epidemiologic studies. *Statistics in Medicine*, 4 : 457 - 467.

Rosner, B., and Munoz, A. (1988). Autoregressive modelling for the analysis of longitudinal data with unequally space examinations. *Statistics in Medicine*, 7 : 59 - 71.

Salvaggio JE, Taylor, G., Weill, H. Occupational asthma and rhinitis. (1986). In: Merchant, J.A., ed. *Occupational Respiratory Diseases*. Washington: United States Department of Health and Human Services. 461-477.

SAS Guide to Macro Processing. SAS Institute Inc., SAS Campus Drive, Cary, North Carolina.

SAS Macro - PHRWLW : Analysis of Multiple Failure Times, SAS Statistics Sample Library, SAS Institute Inc.

SAS Technical Report P-229. (1992). SAS/STAT software: Changes and Enhancements, Release 6.07, Cary, NC: SAS Institute Inc.

Schwartz, D.A., Thorne, P.S., Yagla, S.J., Burmeister, L.F., Olenchock, S.A., Watt, J.L., and Quinn, T.J. (1995). The role of endotoxin in grain dust-induced lung disease. *American Journal of Respiratory and Critical Care Medicine*, 152 : 603 - 608.

Searle, S.R., Casella, G., and McCulloch, C.E. (1992). *Variance components*. John Wiley and sons, New York.

Senthilselvan, A., Chen, Y., and Dosman, J.A. (1993). Predictors of asthma and wheezing in adults: Grain farming, sex, and farming. *American Review of Respiratory Diseases*, 148 : 667 - 670.

Senthilselvan, A., Pahwa, P., Wang, P., McDuffie, H.H., Dosman, J.A. (1996). Persistent wheeze in grain elevator workers should not be ignored. *American Review of Respiratory and Critical Care Medicine*, 153 : 701-705.

Sestini, P., Refini, M.G., Pieroni, P.M., Bracciale, E., Renzoni, F., Tarantino, M., and Vagliasindi, M. (1995). Use of Kaplan-Meier method for he comparison of allergen-specific bronchial reactivity in different groups of patients allergic to grass pollen. *European Respiratory Journal*, 8 (Suppl 19): p1165 - 228s.

Sestini, P., Refini, M.G., Pieroni, P.M., Tarantino, M., Renzoni, F., Vaghi, A., Robuschi, M., and Vagliasindi, M. (1996). Use of Kaplan-Meier method for the statistical analysis of allergen-specific bronchial reactivity: comparison with the slope of the dose-response curve. *American Journal of Respiratory and Critical Care Medicine*, 153 - A862.

Sherrill, D.L., Lebowitz, M.D., Kundson, R.J., and Burrows, B. (1992). Continuous longitudinal regression equations for pulmonary function measures. *European Respiratory Journal*, 5 : 452 - 462.

Sherrill, D.L., Lebowitz, M.D., Kundson, R.J., and Burrows, B. (1993). Longitudinal methods for describing the relationship between pulmonary function, respiratory symptoms and smoking in elderly subjects: The Tucson Study. *European Respiratory Journal*, 6 : 342 - 348.

Sherrill, D.L., and Viegi, G. (1996). On modeling longitudinal pulmonary function data. *American Journal of Respiratory and Critical Care Medicine*, 154 : S217 - S222.

Sparrow, D., O'Connor, G., Colton, T., Barry, C.L., Weiss, S.T., (1987). The relationship of nonspecific bronchial responsiveness to the occurrence of respiratory symptoms and decreased levels of pulmonary function. The Normative Aging Study. *American Review of Respiratory Diseases*, 135 : 1255-1260.

Sparrow, D., O'Connor, G.T,, Basner, R.C., Rosner, B., and Weiss, S.T. (1993). Predictors of the new onset of wheezing among middle-aged and older men. The Normative Aging Study. *American Review of Respiratory Diseases*, 147 : 367 - 371.

Spooner, G.R., Deasi, H.B., Angel, J.F., Reeder, B.A., and Donat, J.R. (1993). Using pyridoxine to treat Carpel Tunnel Syndrome. - Randomized control trial. *Canadian Family Physician*, 39 : 2122 - 2127.

Tabona, M., Chan-Yeung, M., Enarson, D., MacLean, L., Dorken, E., Schulzer, M. (1984). Host factors affecting longitudinal decline in lung spirometry among grain elevator workers. *Chest*, 85 : 782-786.

Therneau, T. M. (1997). Extending the Cox model. *Proceedings of the First Seattle Symposium in Biostatistics*. New York: Springer-Verlag.

Trigg, C.J., Bennett, J.B., Tooley, M., Sibbald, B., D'Souza, M.F., Davies, R.J. A general Practice based survey of bronchial hyperresponsiveness and its relation to symptoms, sex, age, atopy, and smoking. *Thorax*, 45 : 866-872.

Tsiatis, A.A., DeGruttola, V., and Wulfsohn, M.S. (1995). Modeling the relationship of survival to longitudinal data measured with error. Applications to survival and CD4 counts in patients with AIDS. *Journal of the American Statistical Association*, 90 : 27 - 37.

Verberne, A.P.H., Hop, W.C.J., Bos, A.B., and Kerrebijn, K.F. (1993). Effect of a single dose of inhaled salmeterol on baseline airway caliber and methacholine-induced airway obstruction in asthmatic children. *Journal of Allergy and Clinical Immunology*, 127-134.

Vonesh, E.F. (1992). Non-linear models for the analysis of longitudinal data. *Statistics in Medicne*, 11 : 1929 - 1954.

Vonesh, E.F., Chinchilli, V.M., and Pu, K. (1996). Goodness-of-fit in generalized nonlinear mixed-effects models. *Biometrics*, 52 : 572 - 587.

Vonesh, E.F. and Chinchilli, V.M. (1997). *Linear and non-linear models for the analysis of repeated measurements.* Marcel Dekker Inc., New York.

Ware, J.H. (1985). Linear models for the analysis of longitudinal studies. *The American Statistician*, 39 : 95 -101.

Ware, J.H., Dockery, D., Louis, T.A., Xu, X.P., Ferris, B.G.Jr., and Speizer, F.E. (1990). Longitudinal and cross-sectional estimates of pulmonary function decline in never-smoking adults. *American Journal of Epidemiology*, 32 : 685 - 700.

Warren, C.P.W. (1990). Overview of respiratory health risks in agriculture. In: *Principles of Health and Safety in Agriculture*, Ed, Cockcroft, D.W. and Dosman, J.A., pp. 47 - 49. CRC Press, Boca Raton, Florida.

Warren, C.P.W., and Manfreda J. (1980). Respiratory symptoms in Manitoba farmers: association with grain and hay handling. *Canadian Medical Association Journal*, 122: 1259 - 1264.

Wedderburn, R.W.M. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. *Biometrics*, 61 : 439 - 447.

Wei, L.J., Lin, D.Y., and Weissfeld, L. (1989). Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. *Journal of the American Statistical Association*, 84 : 1065 - 1073.

Weiss, S.T., and Speizer, F.E. (1984). Increased levels of airway responsiveness at risk factor for development of chronic lung disease. *Chest*, 86 : 3 - 4.

Winer, B.J. (1971). *Statistical Principles in Experimental Design, 2nd edn.* McGraw-Hill, New York.

Wishart, J. (1938). Growth-rate determination in nutrition studies with the bacon pig and their analysis. *Biometrika*, 30 : 16 - 28.

Wolfinger, R.D., Tobias, R.D., and Sall J. (1991). Structuring Calculations for mixed linear models. *Proceedings of the Statistical Computing Section, American Statistical Association*. 150-155.

Woolcock, A.J., Peat J.K., Salome, C.M., Yan K., Anderson, S.D., Schoeffel, R.E., McCowage, G., and Killalea, T. (1987). Prevalence of bronchial hyperresponsiveness and asthma in rural adult population. *Thorax*, 42 : 361 - 369.

Wulfsohn, M.S., and Tsiatis, A.A. (1997). A joint model for survival and longitudinal data measured with error. *Biometrics*, 53 : 330 - 339.

Zeger, S.L., and Liang, K-Y. (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics*, 42 : 121 - 130.

Zeger, S.L, Liang, K-Y., and Albert P.S. (1988). Models for longitudinal data: A generalized estimating equation approach. *Biometrics*, 44 : 1049 -1060.

Zeger, S.L., Liang, K-Y., and Qaqish B. (1992). Multivariate analyses for categorical data. *Journal of the Royal Statistical Society B*, 54 : 3 - 40.

Zeger, S.L., and Liang K-Y. (1992). An overview of methods for the analysis of longitudinal data. *Statistics in Medicine*, 11 : 1825 - 1839.

Zejda, J.E, Pahwa, P., and Dosman, J.A. (1992). Decline in spirometric variables in grain workers from start of employment: differential effect of duration of follow up. *British Journal of Industrial Medicine*, 49 : 576 - 580.

12. APPENDICES

12.1 APPENDIX A

12.1.1 Definitions and properties of matrices used in the thesis

12.1.2 Basic Asymptotic Results Used in Thesis

12.1.3 Multivariate Gauss-Markoff theorem

12.1.5 Algorithms

12.1.1 Definitions and properties of matrices used in the thesis

Transpose of a Matrix:

An mxn matrix $\underline{A}$ can be written as:

$$\underline{A} = ((a_{ij}))_{i=1,\ldots,m;\ j=1,\ldots,n}$$

The subscripts i and j represents the rows and columns of matrix $\underline{A}$. The transpose of an mxn matrix $\underline{A}$, denoted by $\underline{A}'$, is obtained by simply interchanging the rows and columns of $\underline{A}$ so that $\underline{A}' = ((a_{ij}))_{i'=1,\ldots,n;\ j'=1,\ldots,m}$

Null, Unit and Identity Matrices :

A null matrix, denoted by $\underline{0}$, is one with each element equal to zero.

$$\underline{0} = \begin{bmatrix} 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix}$$

A square matrix with elements all equal to one is referred to as a unit matrix of one's and is denoted by $\underline{J}_p$ (or simply J when the order is clear). Similarly, a unit vector of one's is denoted by $\underline{j}_p$ or $\underline{j}$.

$$\underline{J}_p = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 1 & 1 & \cdots & 1 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & 1 & 1 \end{bmatrix} \text{ and } \underline{j}_p = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}$$

The identity matrix $\underline{I}_p$ is the pxp matrix with one's along its main diagonal and zeros on its off diagonals, i.e.

$$I_p = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}$$

Rank of a Matrix:

Let $\underline{A} = [\underline{a}_1, \underline{a}_2, \ldots, \underline{a}_n]$ be a mxn matrix. The rank of $\underline{A}$ is defined as the number of linearly independent column vectors $\underline{a}_1, \underline{a}_2, \ldots, \underline{a}_n$, of $\underline{A}$. A set of column vectors $\underline{a}_1, \underline{a}_2, \ldots, \underline{a}_n$ is said to be linearly independent if there does not exit a set of scalar coefficient, $c_1, c_2, \ldots, c_n$, (not all zero) such that $c_1a_1 + c_2a_2 + \ldots + c_na_n = 0$. Properties related to rank of a matrix $\underline{A}$ can be found in (Dorf, 1969).

Inverse of a Matrix:

Let $\underline{A}$ be an mxm square matrix with rank($\underline{A}$) = m, then $\underline{A}$ is said to be nonsingular. For any nonsingular matrix $\underline{A}$, there exists a unique inverse $\underline{A}^{-1}$ such that $\underline{A}\,\underline{A}^{-1} = \underline{A}^{-1}\,\underline{A} = \underline{I}$. Some useful results pertaining to inverses can be found in Dorf (1969).

Determinants:

Let $\underline{A}$ be a square matrix of order m. Choose a fixed row value i. The determinant is calculated emanating from the i^{th} row. Let $|\underline{A}|$ denotes the determinant of a square matrix of order m. Then,

$$|\underline{A}| = A_{i,1} \cdot a_{i,1} + A_{i,2} \cdot a_{i,2} + A_{i,3} \cdot a_{i,3} + \ldots A_{i,n} a_{i,n}$$

$A_{i,j}$ is called the cofactor of $a_{i,j}$. The cofactor $A_{i,j}$ is independent of the elements of the i^{th} row and the elements of the j^{th} column.

The value of $A_{i,j} = (-1)^{i+j}$.(the determinant of the sub-matrix of $\underline{A}$, obtained from $\underline{A}$ by crossing out the i^{th} row and the j^{th} column).

The matrix $\underline{A}$ is nonsingular if and only if $|\underline{A}| \neq 0$ or equivalently, if and only if rank($\underline{A}$) = m. Some very important properties of determinants used quite often in this thesis are:

1. For square matrices $\underline{A}$ and $\underline{B}$, $|\underline{A}\,\underline{B}| = |\underline{A}|\,|\underline{B}|$.
2. IF $|\underline{A}|$ is a diagonal or triangular square matrix of order m, $|\underline{A}| = \prod_{i=1}^{m} a_{ii}$
3. For any square matrix of order m, $|a\underline{A}| = a^m |\underline{A}|$, where a is any scalar.
4. $|\underline{A}^{-1}| = 1/|\underline{A}|$

Some other properties of $|\underline{A}|$ can be found in Dorf (1969).

Generalized Inverse:

In many applications, a matrix $\underline{A}$ may be singular (i.e. $|\underline{A}| = 0$) or it may not even be square matrix. In such cases a regular inverse can not be defined and we have recourse to use a generalized inverse (g-inverse) of $\underline{A}$, denoted by $\underline{A}^{-}$. A g-inverse is any matrix which satisfies the following property:

$$\underline{A}\,\underline{A}^{-}\,\underline{A} = \underline{A}$$

When $\underline{A}$ is a square matrix and nonsingular, then $\underline{A}^{-} = \underline{A}^{-1}$ is the unique inverse of $\underline{A}$.

There are infinite number of generalized inverses of $\underline{A}$ each satisfying the above

property. The Moore-Penrose generalized inverse is a unique g-inverse which have additional properties (Rao and Mitra, 1971).

Trace of a Matrix:

The trace of a square matrix $\underline{A}$ of order m is defined to be the sum of the diagonal elements of A and is denoted by trace($\underline{A}$) = $\sum_{i=1}^{m} a_{ii}$ .Some properties involving trace are given in Rao (1973).

Quadratic Forms and Positive Definite Matrix:

A square matrix $\underline{A}$ is said to be symmetric if $\underline{A} = \underline{A}'$. Let $\underline{x} = [x_1 \ldots x_m]'$ be a mx1 vector of variables. $\underline{A}$ quadratic form in $\underline{x}$ is a homogenous quadratic function of $x_1, x_2, \ldots, x_m$ of the form

$$Q = \sum_{i=1}^{m}\sum_{j=1}^{m} a_{ij} x_i x_j = \underline{x}' \underline{A} \underline{x}$$

where $\underline{A} = ((a_{ij}))$ is a symmetric matrix of order m. A real quadratic form $\underline{x}' \underline{A} \underline{x}$ is said to be i) positive definite (p.d.) if $\underline{x}' \underline{A} \underline{x} > 0$ for all non-null $\underline{x}$, ii) positive semidefinite (p.s.d) if $\underline{x}' \underline{A} \underline{x} \geq 0$ for all non-null $\underline{x}$, iii) negative definite (n.d) if $-\underline{x}' \underline{A} \underline{x}$ is p.d., and iv) negative semi-definite (n.s.d) if $-\underline{x}' \underline{A} \underline{x}$ is p.s.d. If $\underline{x}' \underline{A} \underline{x}$ is p.d. or p.s.d., it is also referred as non-negative definite (n.n.d). The matrix $\underline{A}$ of quadratic form $\underline{x}' \underline{A} \underline{x}$ is referred as p.d., p.s.d., n.d., n.s.d., or n.n.d according to the classification of the quadratic form. If $\underline{A}$ is positive definite, then $\underline{A}$ is both symmetric and nonsingular.

Eigenvalues:

For a real symmetric matrix A of order m, the characteristic equation of $\underline{A}$ is given by:

$$|\underline{A} - \lambda \underline{I}| = 0$$

Roots $\lambda_1, \lambda_2, \ldots, \lambda_m$ of above equation are called eigen values.

Direct Product of Matrices:

The direct product of mxn matrix $\underline{A}$ and pxq matrix $\underline{B}$ is the mpxnq matrix defined by:

$$\underline{A} \otimes \underline{B} = ((a_{ij}\ \underline{B}))_{i=1,2,\ldots,\mu;\ j=1,2,\ldots,\nu} = \begin{bmatrix} a_{11}\ \underline{B} & a_{12}\ \underline{B} & \cdots & a_{1\nu}\ \underline{B} \\ a_{21}\ \underline{B} & a_{22}\ \underline{B} & \cdots & a_{2\nu}\ \underline{B} \\ \vdots & \vdots & \ddots & \vdots \\ a_{\mu 1}\ \underline{B} & a_{\mu 2}\ \underline{B} & \cdots & a_{\mu\nu}\ \underline{B} \end{bmatrix}$$

The direct product is also known as the Kronecker product. Kronecker product has several property which can be found in Rao (1973).

Vec(.) and Vech(.) operators:

The Vec(.) operator creates a column vector from a matrix $\underline{A}$ by simply stacking the column vectors of $\underline{A}$ below one another. Hence, for mxn matrix $\underline{A} = [\underline{a}_1, \ldots, \underline{a}_n]$ with column vectors $\underline{a}_1, \ldots, \underline{a}_n$

$$\text{Vec}(\underline{A}) = \begin{bmatrix} \underline{a}_1 \\ \underline{a}_2 \\ \vdots \\ \underline{a}_n \end{bmatrix}$$

Similarly, the Vech(.) operator creates a column vector from a symmetric matrix by stacking the lower diagonal elements below one another. For example;

$$\text{Vech}(A) = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} = \begin{bmatrix} a_{11} \\ a_{21} \\ a_{22} \\ a_{31} \\ a_{32} \\ a_{33} \end{bmatrix}$$

Several properties of Vec(.) and Veach(.) can be found in Vonesh and Chinchilli (1996).

Matrix Differentiation:

Let $\lambda(\underline{x})$ denotes a scalar function of a vector variable $\underline{x} = [x_1, \ldots, x_p]'$, let $\underline{f}(\underline{x}) = [f_1(x), \ldots, f_m(x)]'$ be a mx1 vector function of x and let $\underline{F}(\underline{x}) = [\underline{f}_1(\underline{x}), \underline{f}_2(\underline{x}), \ldots, \underline{f}_n(\underline{x})]$ be a *mxn* matrix function of $\underline{x}$ where each $\underline{f}_i(\underline{x})$ is an mx1 vector function of $\underline{x}$ = *[i = 1, 2, ... , n)*. The gradient or first order derivative vector of $\lambda(\underline{x})$ with respect to $\underline{x}$ is defined to be the *px1* vector

$$\frac{\partial \lambda(\underline{x})}{\lambda(\underline{x})} = \begin{bmatrix} \frac{\partial \lambda(x_1)}{\lambda(x_1)} \\ \vdots \\ \frac{\partial \lambda(x_p)}{\lambda(x_p)} \end{bmatrix}$$

The Hessian or second-order derivative matrix of $\lambda(\underline{x})$ is defined to be the *pxp* matrix :

$$\frac{\partial^2 \lambda(\underline{x})}{\partial \underline{x} \partial \underline{x}'} = \begin{bmatrix} \frac{\partial^2 \lambda(\underline{x})}{\partial x_1^2} & \frac{\partial^2 \lambda(\underline{x})}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 \lambda(\underline{x})}{\partial x_1 \partial x_p} \\ \frac{\partial^2 \lambda(\underline{x})}{\partial x_2 \partial x_1} & \frac{\partial^2 \lambda(\underline{x})}{\partial x_2^2} & \cdots & \frac{\partial^2 \lambda(\underline{x})}{\partial x_2 \partial x_p} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 \lambda(\underline{x})}{\partial x_p \partial x_1} & \frac{\partial^2 \lambda(\underline{x})}{\partial x_p \partial x_2} & \cdots & \frac{\partial^2 \lambda(\underline{x})}{\partial x_p^2} \end{bmatrix}$$

Jacobian Matrix

The Jacobian or first-order derivative matrix of $f(\underline{x})$ is mxp matrix of partial derivatives:

$$\partial \underline{f}(\underline{x})/\partial \underline{x}' = \begin{bmatrix} \partial f_1(\underline{x})/\partial x_1 & \partial f_1(\underline{x})/\partial x_2 & \cdots & \partial f_1(\underline{x})/\partial x_p \\ \partial f_2(\underline{x})/\partial x_p & \partial f_2(\underline{x})/\partial x_p & \cdots & \partial f_2(\underline{x})/\partial x_p \\ \vdots & \vdots & \cdots & \vdots \\ \partial f_m(\underline{x})/\partial x_p & \partial f_m(\underline{x})/\partial x_p & \cdots & \partial f_m(\underline{x})/\partial x_p \end{bmatrix}$$

12.1.2 Basic Asymptotic Results Used in Thesis

$O_p(1)$: If $\{Y_n\}$ is a sequence of random variables with distribution functions $\{D_{Y_n}\}$, then Y_n is said to converge in distribution to the random variable Y having distribution function D_Y, written as : $Y_n \xrightarrow{d} Y$ if

$$\lim_{n \to \infty} D_{Y_n}(y) = D_Y(y)$$

for all y at which $D_Y(y)$ is continuous. If $Y_n \xrightarrow{d} Y$, then $Y_n = O_p(1)$, i.e $\{Y_n\}$ is bounded in probability by 1. [In particular, for a sequence of random variable $\{Y_n\}$ and a

sequence of positive real numbers $\{a_n\}$, then Y_n is bounded in probability by a_n and write: $Y_n = O_p(a_n)$]

$\sqrt{n}$-consistent : In many cases, whether or not a particular estimate is consistent can be shown by simply knowing its limiting distribution. For example, if

$\sqrt{n}(\hat{\underline{\theta}}_n - \underline{\theta}) \xrightarrow{d} N(0, \Omega)$ then $\hat{\theta}_n - \theta = O_p(n^{-1/2}) = o_p(1)$ or $n^{-1/2}(\hat{\theta}_n - \theta) = O_p(1)$ and $\hat{\theta}_n$ is said to be $\sqrt{n}$-consistent.

12.1.3 Multivariate Gauss-Markoff Theorem:

Let $\underline{Y} = \underline{X}\underline{B} + \underline{\varepsilon}$ such that $E(\underline{Y}) = \underline{X}\underline{B}$ and the $V(\underline{Y}) = \underline{I} \otimes \underline{\Sigma}$. If $\hat{\underline{\psi}} = \underline{c}' \underline{B}\underline{a}$ is estimable and $\hat{\underline{B}} = (\underline{X}' \underline{X})^{-} \underline{X}' \underline{Y}$ is a least squares estimate of $\underline{B}$. Then.

(i) $\hat{\underline{\psi}} = \underline{c}' \underline{B}\underline{a}$ is the unique linear unbiased estimate of ψ and has minimum variance among all linear unbiased estimates (that is $\underline{\psi}$ is a best linear unbiased estimate);

(ii) $V(\hat{\psi}) = a' \underline{\Sigma}\, a\, [\underline{c}'(\underline{X}'\underline{X})^{-1}\underline{c}]$; and

(iii) An unbiased estimate of $\underline{\Sigma}$ is given by: $S = \dfrac{(\underline{Y} - \underline{X}\hat{\underline{B}})' (\underline{Y} - \underline{X}\hat{\underline{B}})}{N - r}$

12.1.4 Algorithms

The maximum likelihood is a general method for estimating parameters. Let $L(\underline{y}; \underline{\theta})$ be the density (discrete or continuous) which represents the joint distribution of data $\underline{y}$, and $\underline{\theta}$ be the vector of parameters. The possible values of $\underline{\theta}$ form the parameter space. The density L, a function of $\underline{\theta}$ and for a given data $\underline{y}$ is known as the likelihood function. The maximum likelihood estimator of $\underline{\theta}$ is the value(s) of $\underline{\theta}$, for which the

likelihood function L attains its maximum. Maximum likelihood estimator may not exists, and when it does exist, it may not be unique. It is easier to work with log-likelihood, i.e; $l(\underline{y}; \underline{\theta}) = log\ L(\underline{y}; \underline{\theta})$. The vector of first-order partial derivatives of the $l(\underline{y}; \underline{\theta})$ is called a scoring vector. A standard approach to maximum likelihood estimation is to finding all the roots of the scoring vector and to explore the behaviour of log-likelihood on the boundary of the parameters space and at the points where scoring vector is not defined. Let $l(\underline{y}; \underline{\theta})$ be twice continuously differentiable function of the parameter $\underline{\theta}$. The matrix of second-order partial derivatives is negative definite for each $\underline{\theta}$ and parameter space is an open set. General algorithms for computing Maximum likelihood estimates of the regression and covariance parameters are: Newton-Raphson and Fisher scoring. These algorithms require evaluation of the scoring function $\partial l/\partial\theta_i$ and of either the second-order partial derivatives, $\partial^2 l/(\partial\theta_1\partial\theta_2)$, or the information matrix $[-E\{\partial^2 l/(\partial\theta_1\partial\theta_2)\}]$ for each regression and variance parameters or their pairs, respectively.

NEWTON-RAPHSON METHOD:

The Newton-Raphson algorithm is a general method for finding maximum likelihood estimates. Suppose we have a 'current' solution $\hat{\theta}_{old}$, then the updated solution is defined as:

$$\theta_{new} = \hat{\theta}_{old} - (\partial^2 l/(\partial\theta\partial\theta')^{-1} (\partial l/(\partial\theta) \qquad \text{... (A-1)}$$

where all the partial derivatives are calculated at $\theta = \hat{\theta}_{old}$. The Newton-Raphson algorithm consists of iterations of *(A-1)*, with the 'new' solution from one

iteration becoming the 'old' solution of the next one. The Newton-Raphson method converge to the unique maximum likelihood function under certain regularity conditions. The Newton-Raphson algorithm can be applied to situations where the maximized function is not defined in an open parameter space, or the function's second-order partial derivative matrix is not negative definite throughout the parameter space. In such cases, even if Newton-Raphson iterations converge to a point, there is no assurance that this point is a global maximum. In many situations the matrix of second-order partial derivatives is negative definite and the likelihood has a unique maximum at the unique root of the scoring function. A problem arises when the parameter space is not open, and a solution lies on the boundary. There are few approaches which can be used to handle such a problem (Longford, 1993).

The Best Linear Unbiased Estimator (B.L.U.E):

In mixed model our interest is in linear unbiased estimation of $\underline{\beta}$ or any linear functions of $\underline{\beta}$, say $\underline{k}'\underline{\beta}$, i.e the estimator has the form $\underline{k}'\underline{y}$, and $E(\underline{k}'\underline{y}) = \underline{k}'\underline{\beta}$, if possible. If $\underline{k}'\underline{\beta}$ can be estimated unbiasedly, it is called estimable. Different methods for verifying estimability are given in Henderson's book (1984). If $\underline{k}'\underline{\beta}$ is estimable, then there exists some $\underline{a}$ such that $E(\underline{a}'\underline{y}) = \underline{k}'\underline{\beta}$. Assuming that more than one $\underline{a}$ gives an unbiased estimator, then the one with the minimum sampling variance is chosen, and such an estimator is called the best linear unbiased estimator (B.L.U.E.).

The Best Linear Unbiased Prediction (B.L.U.P):

The estimation of fixed effects vector $\underline{\beta}$ leads to B.L.U.E and prediction of random effects vector $\underline{\gamma}$ leads to best linear unbiased prediction (B.L.U.P). Suppose the predictand is the random variable, $\underline{\gamma}$, and all we know is that it has mean $\underline{k}'\underline{\beta}$, and

variance = $\underline{v}$, and its covariance with $\underline{y}$ is $\underline{c}'$. One possibility to estimate $\underline{\gamma}$ is to find some linear function of $\underline{y}$ that has expectation $\underline{k}'\underline{\beta}$, and in the class of such predictors has minimum variance of prediction errors. This method is called the best linear unbiased prediction (B.L.U.P). Henderson (1984) gave a method how it is derived.

The Multivariate Normal Distribution:

Let $\underline{b}$ be a nonnull px1 vector. A p-variate random vector $\underline{Y}$ is said to have a p-variate normal distribution if $\underline{b}'\underline{Y}$ has a univariate normal distribution for every nonnull px1 vector $\underline{b}$. A p-variate normal distribution of $\underline{Y}$, is uniquely determined by a mean vector $\underline{\mu}$ and covariance matrix $\underline{\Sigma}$ and expressed as

$$\underline{Y} \sim N_p(\underline{\mu}, \underline{\Sigma})$$

Anderson (1984) explains some important results related to multivariate normal.

The Wishart Distribution:

Let $\underline{Y}_1, \underline{Y}_2, \ldots, \underline{Y}_n$ be independently identically distributed random vectors, each with distribution $N_p(\underline{0}, \underline{\Sigma})$. These random vectors can be written as nxp matrix $\underline{Y} = [Y_1, Y_2, \ldots, Y_n]'$. Let

$$Vec(\underline{Y}) \sim N_{np}(\underline{0}, \underline{\Sigma} \otimes \underline{I}) \text{ and}$$

$\underline{Q}_Y = \underline{Y}'\underline{Y} = \sum_{i=1}^{n} Y_i Y_i'$ be the pxp uncorrected sums of squares and crossproducts matrix. Then $\underline{Q}_Y$ is said to have a central Wishart distribution with n degrees of freedom and parameter matrix $\underline{\Sigma}$ and is expressed as:

$\underline{Q}_Y \sim W(n, \underline{\Sigma})$. Some properties and results of central Wishart distribution are given in Anderson (1984).

12.2. APPENDIX B

12.2.1 Definitions of terms related to lung functions measurements

12.2.1 Definitions of terms related to lung functions measurements:

FEV_1 (Forced Expiratory Volume in one second): The FEV_1 is the most important spirometry variable. The FEV_1 is the volume of air exhaled in one second by a forced expiration from full inspiration. The FEV_1 is reduced with airflow limitation or obstruction. The FEV_1 is used to quantify the degree or severity of obstruction.

FVC (Forced Vital Capacity): Spirometry provides a measure of the FVC. The FVC is the volume of air that can be exhaled after the patients takes as deep a breath as possible. The FVC maneuver, a maximal expiratory effort beginning at total lung capacity (TLC) and ending at residual volume is the basis for most spirometry tests.

FEV_1/FVC Ratio: The ratio of the FEV1 to the FVC is used as an indicator of the presence of airflow obstruction. This ratio is used only to indicate the presence of obstruction, not to quantify the severity of obstruction.

$FEF_{25-75\%}$ /MMEF (Forced Expiratory Flow/Maximal Mid-Expiratory Flow from 25% and 75% of the FVC). It is the average flow between 25% and 75% of the exhaled vital capacity and can be measured directly from a volume-time spirogram. It is sensitive index of mild airways obstruction.

Airways responsiveness: Airway responsiveness may be defined as the normal tendency for airways to constrict under the influence of various stimuli.

Hyperreactivity: is a term used to describe airways which narrow excessively in response to various inhaled agents. Usually measured by the decrease in FEV_1 following inhalation of methacholine. Almost all asthmatics have hyperreactive airways.

Obstruction: is a disease in maximal airflow rates caused by airway narrowing.

PC_{20} : Bronchial responsiveness is evaluated by the provocation test. A positive test results when inhalation of stimuli produces a 20% decrease in the FEV_1. The values for PC_{20} are interpolated from a log concentration-response curve. A 20% fall in FEV_1 is the most commonly used clinical response. The concentration of stimuli at which the 20% decrease occurs is referred to as the provocation concentration (PC_{20}), which is used as an indicator of bronchial hyperresponsivenes

12.3 APPENDIX C

12.3.1. Medical Practitioner's Report to Labour Canada

12.3.2 Grain Dust Questionnaire

12.3.3 Prick Skin Test

MEDICAL PRACTITIONER'S REPORT TO LABOUR CANADA (MODIFIED)

SENIOR CONSULTANT, OCCUPATIONAL MEDICINE
OCCUPATIONAL SAFETY AND HEALTH BRANCH
LABOUR CANADA, OTTAWA, ONTARIO K1A 0J2

NOTE: SHADED AREAS ARE FOR LABOUR CANADA USE ONLY

R ☐ P ☐

1. IDENTIFICATION

EMPLOYEE CODE NO.	MALE	FEMALE	AGE	NO. OF YEARS IN GRAIN INDUSTRY

COMPANY

TYPE OF ELEVATOR ☐ PRIMARY (COUNTRY) ☐ TERMINAL (TRANSFER) ☐ PROCESS

ELEVATOR NAME/LOCATION (EXCEPT PRIMARY)

JOB TITLE

ELEVATOR WORK AREA (EXCEPT PRIMARY)

EXAMINATION ☐ PREPLACEMENT ☐ SPECIAL (SPECIFY) _____ PERIODIC ☐ 1st ☐ 2nd ☐ 3rd ☐ 4th ☐ 5th

DATE OF EXAMINATION

GOC ☐

2. SMOKING ☐ NON-SMOKER

SMOKED PREVIOUSLY: ___ YEARS; DATE STOPPED _____
NO. OF CIGARETTES DAILY _____
NO. OF CIGARS DAILY _____
NO. OF PIPEFULLS DAILY _____

SMOKES: FOR ___ YEARS
NO. OF CIGARETTES DAILY _____
NO. OF CIGARS DAILY _____
NO. OF PIPEFULLS DAILY _____

P1 ☐ P2 ☐

3. CHEST X-RAY RESULTS

☐ NORMAL ☐ ABNORMAL — BRIEF DESCRIPTION _____

☐ ☐

4. PULMONARY FUNCTION RESULTS

FVC (LITRES)	FEV_1 (LITRES)	FEV_1/FVC X 100	MMEF (LITRES/SEC)

HEIGHT _____ cm
WEIGHT _____ kg

5. CLASSIFICATION OF OBSERVED RESPIRATORY FUNCTION (SEE REVERSE SIDE OF FORM FOR GUIDELINES)

	YES	NO
CHRONIC COUGH	☐	☐
CHRONIC SPUTUM	☐	☐
WHEEZE	☐	☐
DYSPNEA ON EFFORT	☐	☐

☐ GRADE 1 ☐ GRADE 2 ☐ GRADE 3

GRADE CHANGED SINCE PREVIOUS EXAMINATION: ☐ YES ☐ NO

6. RE-EXAMINATION RECOMMENDED ☐ NO ☐ YES ON _____ (DATE)

7. COMMENTS _____

MEDICAL PRACTITIONER (NAME — PLEASE PRINT)	(ADDRESS — PLEASE PRINT)
SIGNATURE	TELEPHONE

W ☐ X ☐ Y ☐ Z ☐

FORM 271-4

Symptoms: To Be Recorded

(1) Chronic Cough	YES/NO
(2) Chronic Sputum	YES/NO
(3) Wheeze	YES/NO
(4) Shortness of breath on effort	YES/NO

Definitions:

(1) Cough in the morning or during the day or night for more than three months a year for two years.

(2) Phlegm in the morning or during the day or night for more than three months a year for two years.

(3) Chest wheezing or whistling most days or nights.

(4) Shortness of breath on effort on the level or walking up a slight hill.

Lung Function

Based on forced expiratory spirometry.

Grade (1) FEV_1/FVC x 100 more than 70 per cent

Grade (2) FEV_1/FVC x 100 less than 70, more than 50 per cent

Grade (3) FEV_1/FVC x 100 less than 50 per cent

12.3.2

Study No.:_____

GRAIN DUST QUESTIONNAIRE

Employer: _____

Elevator Location: _____

Height in cm._____ Weight in kg._____ Telephone No. _____

Soc. Ins. No. _____

1. Name: _____
2. Address: _____

_____

3. Date of Birth: _____
4. Age: _____
5. Sex: _____
6. Marital Status: _____
7. Physician's Name and Address: _____

_____

8. Present Occupation: _____
9. Racial Origin: _____
10. Date Studied: Month _____

Day _____

Year _____

11. Time of Study: _____ Holiday? 1. Yes 2. No

Workday? 1. Yes 2. No

12. List all jobs, occupations or types of work you have held or done throughout your life and state approximate dates and lengths of time:

Company or Industry	Job Classification	Type of work or task	Length of time in years	Years From To	Average No. of months per year	Name of location -- City or Rural Area
1.				19___ 19___		
2.				19___ 19___		
3.				19___ 19___		
4.				19___ 19___		
5.				19___ 19___		
6.				19___ 19___		
7.				19___ 19___		
8.				19___ 19___		

Total years exposure to elevator grain dust _____

(CHECK APPROPRIATE ANSWER AFTER EACH QUESTION: WHEN IN DOUBT, ANSWER 'NO')

COUGH AND PHLEGM:

13 a. Do you usually cough first thing in the morning? (Exclude clearing throat) 1. Yes 2. No

b. Do you usually cough at other times during the day or night? 1. Yes 2. No

c. Do you cough as much as 4 - 6 times a day for 4 or more days out of the week? 1. Yes 2. No

IF ANSWER IS "YES" TO EITHER 13 a, b or c, ANSWER d and e:

d. Do you cough on most days for as much as 3 months of the year? 1. Yes 2. No

e. For how many years have you had this cough? _____ years

14 a. Do you usually bring up phlegm from the chest first thing in the morning? (Not from the back of your nose. Count swallowed phlegm from the chest.) 1. Yes 2. No

b. Do you usually bring up phlegm from the chest at other times during the day or night? 1. Yes 2. No

c. Do you bring up phlegm like this as much as twice a day, 4 or more days out of the week? 1. Yes 2. No

IF ANSWER IS "YES" TO EITHER 14 a, b or c, ANSWER d and e:

d. Do you bring up phlegm from the chest on most days for as much as 3 months of the year? 1. Yes 2. No

e. For how many years have you raised phlegm from the chest? _____ years

IF YOU NEVER HAD COUGH OR PHLEGM, GO TO QUESTION 21

15. When is your cough worse?

1. On workdays
2. On weekends when not working
3. I notice no difference

16. Is your cough and/or phlegm better, the same or worse when on vacation or not working?

1. Better
2. The same
3. Worse

17 a. Is your cough and/or phlegm worse at different times of the year?

1. Yes 2. No

If 'No' go to Question 18

IF "YES" TO 17 a, CIRCLE THE MONTHS IN WHICH YOU HAVE BEEN MOST TROUBLED:

b.

January	February	March	April	May	June
1	2	3	4	5	6

July	August	September	October	November	December
7	8	9	10	11	12

18. Is your cough and/or phlegm brought on by, or made worse by exposure to:

a. Grain dust at work? 1. Yes 2. No

b. Other dusts at work? 1. Yes 2. No

c. Gases or fumes at work? 1. Yes 2. No

d. House dust or fumes in the home? 1. Yes 2. No

e. Barn dusts, silage or hay? 1. Yes 2. No

f. Weather changes? 1. Yes 2. No

g. Cold air? 1. Yes 2. No

h. Cigarette smoke? 1. Yes 2. No

i. Other (Please specify) 1. Yes 2. No

IF ANSWER WAS "YES" TO QUESTION 18 a. GRAIN DUST AT WORK?, PLEASE ANSWER QUESTION 19: (Otherwise, go to Question 21 a.)

19. In your opinion, which grain dusts are most likely to bring on cough and/or phlegm, or make it worse? (May circle more than one)

1. Durum wheat
2. Spring wheat
3. Rye
4. Oats
5. Barley
6. Corn
7. Flax
8. Rape
9. Mustard
10. Alfalfa
11. Other _____ (Specify)

20. When you are working regularly, how frequently (on the average) have you experienced cough and/or phlegm during work?

1. Usually at least once a day
2. Only a few times each week
3. Only a few times each month
4. Only a few times each year
5. Only a few times ever
6. Only once

WHEEZING AND/OR CHEST TIGHTNESS

21 a. Have you ever noticed any wheezing and/or tightness in your chest? 1. Yes 2. No

If 'No' go to Question 37

IF ANSWER IS "YES" TO 21 a, ANSWER b and c:

b. Do you get this only with colds? 1. Yes 2. No

c. Do you get this even when you don't have a cold? 1. Yes 2. No

IF YOU HAVE NEVER NOTICED WHEEZING AND/OR TIGHTNESS IN YOUR CHEST, SKIP QUESTIONS 22 to 36, AND GO TO QUESTION 37.

22. Which of these symptoms have you experienced: wheezing, chest tightness or both?

1. Only wheezing
2. Only chest tightness
3. Mainly wheezing
4. Mainly chest tightness
5. Both wheezing and chest tightness

23. At what age did your wheezing and/or chest tightness first occur? _____ years

24. At what age did wheezing and/or chest tightness last occur? _____ years

(If you are still having these, put your present age)

25. Do you have wheezing and/or chest tightness <u>at work</u> while you are performing your job? 1. Yes 2. No

If 'No' go to Question 28

26. When you are working regularly, how frequently (on the average) have you experienced wheezing and/or chest tightness during work?

1. Usually at least once a day
2. Only a few times each week
3. Only a few times each month
4. Only a few times each year
5. Only a few times ever
6. Only once

27. Is your wheezing and/or chest tightness usually <u>worse</u> on:

1. First day back to work
2. Any day(s) at work
3. Weekends, when not working
4. Makes no difference

28. Is your wheezing and/or chest tightness brought on by, or made worse by, exposure to:

a.	Grain dust at work?	1. Yes 2. No
b.	Other dusts at work?	1. Yes 2. No
c.	Gases or fumes at work?	1. Yes 2. No
d.	House dust or fumes in the home?	1. Yes 2. No
e.	Barn dusts, silage or hay?	1. Yes 2. No
f.	Moldy or musty barn dusts, silage or hay?	1. Yes 2. No
g.	Contacts with animals?	1. Yes 2. No
h.	Plants, pollens or weeds?	1. Yes 2. No
i.	Weather changes?	1. Yes 2. No
j.	Cold air?	1. Yes 2. No
k.	Cigarette smoke?	1. Yes 2. No
l.	Other exposures	1. Yes 2. No

_____ (Specify)

IF ANSWER IS "YES" TO QUESTION 28 a, GRAIN DUST AT WORK, PLEASE ANSWER QUESTIONS 29 to 32: (Otherwise, go to Question 33 a.)

29. In your opinion, which grain dusts are most likely to bring on wheezing and/or chest tightness or make it worse? (May circle more than one)

1. Durum wheat
2. Spring wheat
3. Rye
4. Oats
5. Barley
6. Corn
7. Flax
8. Rape
9. Mustard
10. Alfalfa
11. Other

_____ (Specify)

30. When is your wheezing and/or chest tightness <u>most likely</u> to start or get worse? (Circle only one)

1. Before work
2. During work
3. After work
4. Either during or after work

31. If it starts or gets worse during work, how soon after the beginning of the work shift does this happen?

1. Right away

OR

2. _____ hours after

32. If it starts or gets worse after work, how many hours <u>after work</u> does this happen?

_____ hours after

33 a. Does wheezing and/or chest tightness ever wake you up from your sleep?

1. Yes 2. No

If 'No' go to Question 34 a.

IF ANSWER IS "YES" TO QUESTION 33 a, PLEASE ANSWER b.:

b. How often does this happen?

1. Almost every night
2. A few times each month
3. A few times each year
4. A few times ever
5. Only once
6. Never

34 a. Is your wheezing and/or chest tightness worse at different times of the year?

1. Yes 2. No

If 'No' go to Question 35

IF ANSWER IS "YES" TO QUESTION 34 a, CIRCLE THE MONTHS IN WHICH YOU ARE MOST TROUBLED BY WHEEZING AND/OR CHEST TIGHTNESS:

b.

January	February	March	April	May	June
1	2	3	4	5	6

July	August	September	October	November	December
7	8	9	10	11	12

35. Is your wheezing and/or chest tightness better, the same or worse when on vacation or not working?

1. Better
2. The same
3. Worse

36. Have you ever had 2 or more attacks of wheezing that has made you feel short of breath? 1. Yes 2. No

SHORTNESS OF BREATH

37. Have you been troubled by shortness of breath? 1. Yes 2. No

38. Are you troubled by shortness of breath when hurrying on level ground or walking up a light hill? 1. Yes 2. No

39. Do you get short of breath walking with other people of your own age on level ground? 1. Yes 2. No

40. Do you have to stop for breath while walking at your own pace on level ground? 1. Yes 2. No

41. Do you get short of breath dressing or walking about the house? 1. Yes 2. No

42. For how long have you had this shortness of breath? _____ years

43. Do you get short of breath while at work, performing your job? 1. Yes 2. No

44 a. Do you get short of breath during or after exposure to grain dust? 1. Yes 2. No

If 'No' go to Question 45 a.

IF ANSWER IS "YES" TO QUESTION 44a, PLEASE ANSWER QUESTIONS 44 b, c, d and e:

44 b. In your opinion, which grain dusts are most likely to bring on shortness of breath or make it worse? (May circle more than one)

1. Durum wheat
2. Spring wheat
3. Rye
4. Oats
5. Barley
6. Corn
7. Flax
8. Rape
9. Mustard
10. Alfalfa
11. Other

_____ (Specify)

c. When is your shortness of breath most likely to get worse? (Circle only one)

1. During work
2. After work
3. Either during or after work

d. If it starts during work, how soon after the beginning of the work shift does this happen?

1. Right away

OR

2. _____ hours after

e. If it starts after work, how many hours after work?

_____ hours after

IF IN YOUR WORK YOU ARE EXPOSED TO GRAIN DUST, PLEASE ANSWER QUESTIONS 45 to 49, IF NOT, GO TO QUESTION 50:

FEVER AND/OR CHILLS (SHIVERING):

45 a. Have you ever had fever and/or chills during exposure, or after being exposed, to grain dust?

1. Yes 2. No

If 'No' go to Question 48

b. If your answer is "Yes" to Question 45 a, did you have --

1. Only fever
2. Only chills
3. Mostly fever
4. Mostly chills
5. Both fever and chills

46. When have you noticed the fever and/or chills?

1. During work
2. After work
3. Either during or after work

IF IT STARTS AFTER WORK:

47 a. About how many hours after work did this (these) happen? _____ hours after work

b. About how many hours did this (these) last? _____ hours

c. How many times in your work life as a grain handler have you had fever and/or chills after work? _____ times

d. When have you experienced this fever and/or chills?

1. On first day back to work
2. Any other day at work
3. On either the first day back or any other day

e. If on the first day back to work, how long had you been off work? _____ number of days

48. During exposure to grain dust have you ever had:

a. Eyes burning, watering or itching? 1. Yes 2. No

b. Stuffy nose? 1. Yes 2. No

c. Throat sore or burning? 1. Yes 2. No

IF ANSWER IS "YES" TO QUESTIONS 48 a, b or c, PLEASE ANSWER d:

d. In your opinion, which grain dusts are most likely to bring on these symptoms or make them worse? (May circle more than one)

1. Durum wheat
2. Spring wheat
3. Rye
4. Oats
5. Barley
6. Corn
7. Flax
8. Rape
9. Mustard
10. Alfalfa
11. Other _____ (Specify)

49 a. During or immediately after exposure to grain dust, have you ever had itching on your skin?

1. Yes 2. No

If 'No' go to Question 50

IF YOUR ANSWER IS "YES" TO QUESTION 49 a, PLEASE ANSWER b and c:

b. How many times in a year is this likely to happen?

_____ times

c. In your opinion, which grain dusts are most likely to bring on the skin itching? (May circle more than one)

1. Durum wheat
2. Spring wheat
3. Rye
4. Oats
5. Barley
6. Corn
7. Flax
8. Rape
9. Mustard
10. Alfalfa
11. Other

_____
(Specify)

TOBACCO SMOKING

50. Have you ever smoked cigarettes? (If you have smoked less than 20 packs of cigarettes in your lifetime, check 'No')

1. Yes 2. No

If 'No' go to Question 53 a.

51 a. Do you now smoke cigarettes? (Answer "yes" if you currently smoke or if you stopped smoking within the last month)

1. Yes 2. No

If 'No' go to Question 52 a.

IF YOU SMOKE REGULARLY NOW:

b. Do you inhale the cigarette smoke?

1. Yes 2. No

c. How old were you when you began to smoke cigarettes?

_____ years

d. How many cigarettes do you usually smoke each day at the present time? (Please give best estimate: one pack contains 20 cigarettes)

_____cigarettes per day

51 e. What is the usual number of cigarettes you have smoked per day since you began to smoke? (Please give best estimate: one pack contains 20 cigarettes) _____ cigarettes per day

f. If there have been periods when you abstained from smoking, please enter total years of abstinence from smoking. (If less than a year, do not fill in) _____ years

IF YOU HAVE COMPLETED THIS SECTION (QUESTIONS 50 & 51), GO TO QUESTION 53 a.

52 a. Did you used to smoke cigarettes? 1. Yes 2. No

IF YOU DO NOT SMOKE CIGARETTES REGULARLY NOW, BUT USED TO SMOKE THEM: (If you have not smoked at least 20 packs of cigarettes in your lifetime, check here: _____)

b. How old were you when you began to smoke cigarettes? _____ years

c. How old were you when you stopped smoking cigarettes regularly? _____ age

d. What was the usual number of cigarettes you smoked per day? (Please give best estimate: one pack contains 20 cigarettes) _____ cigarettes per day

e. If there have been periods when you abstained from smoking, please enter total years of abstinence from smoking. (If less than year, do not fill in) _____ years

53 a. Do you now smoke pipes or cigars? 1. Yes 2. No

If 'No' go to Question 54

b. Do you usually inhale when you smoke either pipes or cigars? 1. Yes 2. No

PESTICIDES

54. Have you ever been exposed to pesticides?

1. Yes 2. No

If 'No' go to Question 63 a.

55. During or immediately after exposure to pesticides, have you ever had any health problems or symptoms?

1. Yes 2. No

IF YOUR ANSWER IS "YES" TO QUESTION 55, PLEASE ANSWER QUESTIONS 56 to 62:

56. **Where did this (these) exposure(s) happen?**

1. **At work**
2. **At home**
3. **On a farm**

57. **What kind of health problems did you have?**

1. **Weakness**
2. **Fainted**
3. **Dizziness**
4. **Headache**
5. **Convulsions**
6. **Trouble breathing**
7. **Nausea and/or vomiting**
8. **Stomach pain**
9. **Diarrhea**
10. Muscle twitching, cramps
11. Blurred vision
12. Jaundice
13. **Other**

_____
(Specify)

58. How many days did these problems last?

_____ days

59. How many times have you had these problems?

_____ times

60. Have you ever been so ill following the exposure to pesticides that you couldn't do regular work?

1. Yes 2. No

61. Have you ever had to go, or be taken, to a doctor or hospital because of these problems? 1. Yes 2. No

62. What pesticides caused you to have symptoms?

1. Do not know
2. Carbon tet (weevilcide)
3. Malathion
4. Methyl Bromide
5. Phostoxin
6. Other _____ (Specify)

THE NEXT SET OF QUESTIONS ARE ABOUT ILLNESSES YOU HAVE HAD, OR HAVE CURRENTLY. WHEN RECORDING AGE, WRITE IN THE YOUNGEST AGE AT WHICH THE ILLNESS OCCURRED.

63 a. During the past 3 years, how much trouble have you had with illnesses such as chest colds, bronchitis or pneumonia?

1. None
2. Little
3. Moderately
4. Much
5. A great deal

b. During the past 3 years, how often were you unable to do your usual activities because of illnesses such as chest colds, bronchitis or pneumonia?

1. None
2. One time
3. 2 to 5 times
4. More than 5 times

64. Has a doctor ever told you that you had any of the following?

			At age
a. Bronchitis (or bronchial trouble)	1. Yes	2. No	____
b. Emphysema	1. Yes	2. No	____
c. Pleurisy	1. Yes	2. No	____
d. Tuberculosis of the lung	1. Yes	2. No	____
e. Cancer of the lung	1. Yes	2. No	____
f. Chest surgery (including heart surgery)	1. Yes	2. No	____
g. Chest injury	1. Yes	2. No	____
h. Sinus trouble	1. Yes	2. No	____
i. Farmer's Lung disease	1. Yes	2. No	____

65 a. Has a doctor ever said you had:

Pneumonia or broncho-pneumonia? 1. Yes 2. No

If 'No' go to Question 66 a.

IF YOUR ANSWER IS "YES" TO QUESTION 65 a, PLEASE ANSWER b and c:

b. How many times have you had pneumonia? _____ times

c. Your age (or ages) when this (these) happened?

_____, _____, _____, _____ years

66 a. Has a doctor every said you had bronchial asthma?

1. Yes 2. No

If 'No' go to Question 67

IF YOUR ANSWER IS "YES" TO QUESTION 66 a, PLEASE ANSWER b, c and d:

b. How old were you when your asthma started? _____ age started

c. Do you still have asthma? 1. Yes 2. No

d. If no, how old were you when your asthma stopped? _____ age stopped

67. Has a doctor ever told you that you had any of the following?

a. Heart trouble 1. Yes 2. No

b. High blood pressure 1. Yes 2. No

c. Allergic reaction in your nose, such as hay fever 1. Yes 2. No

d. Kidney trouble 1. Yes 2. No

e. Liver trouble or jaundice 1. Yes 2. No

f. Diabetes 1. Yes 2. No

g. Cystic fibrosis 1. Yes 2. No

68. Have you ever had a serious skin rash in infancy? (eczema) ?

1. Yes 2. No

69. Have you ever suffered from skin rashes?

1. Yes 2. No

If 'No' go to Question 71

70 a. If Yes, have you ever suffered from skin rashes lasting longer than 2 weeks?

1. Yes 2. No

If 'No' go to Question 71

IF YOUR ANSWER IS "YES" TO QUESTION 70 a, PLEASE ANSWER b:

b. What area was involved?

1. Face
2. Ears
3. Scalp
4. Hands
5. Arms
6. Chest
7. Back
8. Abdomen
9. Legs
10. Feet

71 a. Have you ever suffered with painful or swollen joints?

1. Yes 2. No

If 'No' go to Question 72

IF YOUR ANSWER IS "YES" TO QUESTION 71 a, PLEASE ANSWER b and c:

b. Which joints were involved?

1. Fingers
2. Wrists
3. Elbows
4. Shoulders
5. Spine
6. Hips
7. Knees
8. Ankles

c. Were the joints swollen?

1. Yes 2. No

72. Do you have frequent "chills" with fever, sweating and perhaps shaking?

1. Yes 2. No

73. Do you have swelling of both ankles?

1. Yes 2. No

74. Has any member of your immediate family (blood relative), had any of the following diseases?

				Relative
a.	Chronic bronchitis	1. Yes	2. No	_____
b.	Emphysema	1. Yes	2. No	_____
c.	Asthma	1. Yes	2. No	_____
d.	Hay fever	1. Yes	2. No	_____
e.	Cystic fibrosis	1. Yes	2. No	_____
f.	Cancer of the lung	1. Yes	2. No	_____
g.	Farmer's Lung disease	1. Yes	2. No	_____
h.	Other lung disease	1. Yes	2. No	_____

_____
(Specify)

75 a. Have you ever had a chest x-ray in the past? 1. Yes 2. No

IF YOUR ANSWER IS "YES" TO QUESTION 75 a, PLEASE ANSWER b and c:

b. Where was the last chest x-ray taken?

_____ (Hospital) in _____ (City - Town) in 19___

OR

_____ (Doctor's office) in _____ (City - Town) in 19___

c. Have you ever been told you had an abnormal chest x-ray? 1. Yes 2. No

76. Are you taking any drugs or medications? (Prescribed or not) 1. Yes 2. No

If yes, please list the medications here:

_____

_____

_____

_____

77. When was the last time you were exposed to your working environment?

1. Today
2. Yesterday
3. 2 days ago
4. _____ days ago

78. Do you live or work on a farm?

1. Yes 2. No

79. If so, for how many years?

_____ years

Date: _____ Signature: _____

Pulmonary Function Survey
April, 1979
Reprinted: March, 1982

GRAIN DUST SURVEY

2.3.3 PRICK SKIN TESTS

NAME:
SUBJECT NO:
DATE:
EMPLOYER:

ROUTINE:

1	CONTROL	
2	ALTERNARIA	
3	ASPERGILLUS	
4	PENICILLIUM	
5	HORMODENDRUM	
6	MITE	
7	MIXED ANIMAL DANDERS	
8	MIXED GRASS POLLEN	
9	MIXED WEED POLLEN	
10	MIXED TREE POLLEN	
11	WHEAT DUST EXTRACT	
12	RYE DUST EXTRACT	
13	OAT DUST EXTRACT	
14	BARLEY DUST EXTRACT	
15	RAPESEED EXTRACT	

LEGEND:

1-2mm +
3-5mm ++
6-8mm +++
8mm ++++